A STRATEGY FOR STEPWISE REGRESSION PROCEDURES IN SURVIVAL ANALYSIS WITH MISSING COVARIATES

by

Jia Li

B.S., Beijing Normal University, 1998

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Jia Li

It was defended on

July 25th 2006

and approved by

Dissertation Advisor:

Stewart J. Anderson, Ph.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Joseph P. Costantino, Dr.PH., Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Jong-Hyeon Jeong, Ph.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Kevin E. Kip , Ph.D., Assistant Professor, Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh

A STRATEGY FOR STEPWISE REGRESSION PROCEDURES IN SURVIVAL ANALYSIS WITH MISSING COVARIATES

Jia Li, PhD

University of Pittsburgh, 2006

The selection of variables used to predict a time to event outcome is a common and important issue when analyzing survival data. This is an essential step in accurately assessing risk factors in medical and public health studies. Ignoring an important variable in a regression model may result in biased and inefficient estimates of the outcomes. Such bias can have major implications in public health studies because it may cause potential risk factors to be falsely declared as associated with an outcome, such as mortality or conversely, be falsely declared not associated with the outcome. Stepwise regression procedures are widely used for model selection. However, they have inherent limitations, and can lead to unreasonable results when there are missing values in the potential covariates. In the first part of this dissertation, multiple imputations are used to deal with missing covariate information. We review two powerful imputation procedures, Multiple Imputation by Chain Equations (MICE) and estimation/multiple imputation for Mixed categorical and continuous data (MIX) that implement different multiple imputation methods. We compare the performance of these two procedures by assessing the bias, efficiency and robustness in several simulation studies using time to event outcomes. Practical limitations and valuable features of these two procedures are also assessed. In the second part of the dissertation, we use imputation together with a criterion called the Brier Score to formulate an overall stepwise model selection strategy. The strategy has the advantage of enabling one to perform model selection and evaluate the predictive accuracy of a selected model at the same time, all while taking into account the missing values in the covariates. This comprehensive strategy is implemented by defining the

Weighted Brier Score (WBS) using weighted survival functions. We use simulations to assess this strategy and further demonstrate its use by analyzing survival data from the National Surgical Adjuvant Breast and Bowel Project (NSABP) Protocol B-06.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

First of all, I would like to thank my advisor, Dr. Stewart J. Anderson, for his guidence, encouragement and support throughout years of my graduate study. I have learned not only sound ways of doing research from him, but also being an honest and caring person. I would also like to thank my committee members Dr. Joseph P. Costantino, Dr. Jong-Hyeon Jeong and Dr. Kevin Kip for their valuable suggestions and thoughtful comments throughout the course of this dissertation.

It has been my pleasure to work in NSABP biostatistics center. I would like to express my appreciation to NSABP and the support of its grant: NIH/NCI 5–U10–CA69651.

Many Thanks to my parents for their years of care and encouragement and to my friends for their continued support.

Finally, I wish to express special thanks and to dedicate this work to my husband, Lijun, for his endless care. I simply could not have taken the inital step of my study nor have gone this far without him. I would like to share my happiness with him.

1.0 INTRODUCTION

1.1 MOTIVATION

Many practical problems involve "time to event" data, and the examples of such data appear in various fields such as clinical trials, cohort studies and epidemiology studies. Often, researchers are interested in comparing different treatment groups. The subjects in the groups may have additional characteristics that could be related and therefore, should be accounted for when a treatment comparison is made. For example, subjects have many recorded demographic variables, e.g., age or gender; physiological variables, e.g., blood pressure or blood glucose levels; behavioral variables, e.g., diet or smoking status. Such variables, which are called independent variables, or covariates, may be used to explain the outcome, response, or dependent variables. A popular method to model this kind of data is Cox proportional hazards model [1]. For example, when comparing two groups, the estimates of parameters of the Cox proportional hazards model will be less biased and more precise than a simple comparison that does not adjust for potential independent variables. However, when some of the potential independent variables have missing observations, standard techniques of estimation may lead to inefficient or even more biased results [2], especially when one tries to select which independent variables have statistically significant effects on the outcome, or determines the best subset of the covariates, and the corresponding best-fitting regression model for indicating the relationship between the outcome and covariates.

There are several model selection regression strategies for survival analysis, and one that is commonly used is the stepwise regression procedure. Stepwise regression was developed in the 1960's. It is the method to choose a subset of independent variables by using only a few possible sub-models. Miller [3] gave a comparison between the stepwise regression

methods and other model selection procedures. Of note, all these methods can be very sensitive to the absence of observations of covariates. When there are missing values among covariates in the model, using a stepwise regression procedure may lead to serious problems, or result in coefficients that have the wrong sign leading to an incorrect inference about the association among the covariates and outcome. One simple way to handle the missing data problem in covariates of Cox proportional hazards model is called complete case analysis, that is, to analyze only those subjects with fully observed data and discard those subjects with unobserved data. However, a complete case analysis may be inefficient and wasteful because part of the information is deleted, especially when the fraction of missingness is large. Moreover, it may also introduce bias when missingness is not completely at random.

Because of these problems, many investigators have proposed methods for employing information regarding missing data. Little and Rubin [4] classified missing data into three categories: the first one is missing completely at random (MCAR), which means that missingness does not depend on the values of missing or observed data; the second one is missing at random (MAR), which means that missingness depends only on the components of observed data and not on the components that are missing; the third one is not missing at random (NMAR), which means that missingness depends on not only the values of observed data, but also the values of unobserved data.

Many approaches have been proposed to improve estimates of parameters of interest in Cox proportional hazards model when there are missing elements in covariates under MCAR or MAR, such as weighted estimating equations methods in Lipsitz, Ibrahim and Zhao [5], Wang and Chen [6] and Parzen, *et al.* [7]; likelihood-based methods in Lin and Ying [8], Lipsitz and Ibrahim [9] and Herring and Ibrahim [10]; an imputation method in Paik and Tsai [11], and multiple imputation methods in Buuren *et al.* [12]. In addition, many researchers have discussed methods to improve stepwise selection procedures [13]. However, the performance of stepwise methods in Cox proportional hazards model with missing covariates has not been investigated. Therefore, in this dissertation, we develop a strategy of using stepwise regression procedures for the Cox proportional hazards model when there is unavailable information among covariates.

This dissertation is organized as two parts. In the first part, which includes chapters

2 and 3, we introduce and evaluate two multiple imputation methods for mixed types of missing covariates in Cox proportional hazards models. In the second parts, which includes chapters 4, 5, 6 and 7, we propose an overall stepwise model selecetion strategy that accounts for loss information due to missing values in covariates.

1.2 A MOTIVATING EXAMPLE

In August 1976, the National Surgical Adjuvant Breast and Bowel Project (NSABP) started the B-06 study, a randomized clinical trial designed to determine whether lumpectomy with or without radiation therapy was as effective as total mastectomy for the treatment of invasive breast cancer [14] [15]. Patient accrual was terminated on January 31, 1984. A total of 2,163 women with invasive breast tumors that were 4 cm or less in their largest diameter and with either negative or positive axillary lymph nodes were entered the study and randomly assigned to one of three treatment groups: total mastectomy (TM), lumpectomy (which is also called segmental mastectomy (SM)), or lumpectomy followed by breast irradiation. Axillary nodes were removed regardless of the treatment assignments.

By the end of 1998, 58 of the 2,163 women did not have follow-up. 81 of the 2,105 follow-up patients were ineligible for study, 165 refused to accept the assigned treatment and 8 patients had no records of nodal status, leaving 1,851 patients available on study. The distribution of women among the treatment groups is shown in Table 1. Table 2 shows that there are moderate missing values in 6 of 8 listed covariates, and 50.35% of women had at least one missing value in their covariates.

A stepwise selection procedure were applied to select important variables from candidate variables for the Cox proportional hazards model. The related statistical analysis indicated that stepwise regression methods in the presence of missing covariates resulted in unstable models, and the selected variables in each model were sensitive to the missing information in the data. This example lead to the work in this dissertation.

Table 1: The Distribution of Women Among The Treatment Groups

Treatment	Frequency	Percent(%)
TM	589	31.82
SM	634	34.25
SM+BI	628	33.93
Total	1851	100

Table 2: Percentage of Missing Values in Covariates

Variable	Percent (%)
Age	0
Race	5.56
Estrogen Receptor	25.23
Progesterone Receptor	37.06
Tumor Type	13.56
Nodal Status	0
Nuclear Grade	10.97
Blood Vessel Invasion	15.40
At Least One Missing	50.35

2.0 LITERATURE REVIEW ON MISSING COVARIATES

Most standard techniques for regression models in survival analysis require full covariate information. However, many covariates are partially unavailable in most clinical trials or observational studies. One simple way to solve the missing data problem is to analyze only the subjects with complete observations. This method is called a complete case analysis. However, the complete case method can lead to biased results and large standard errors when the missing data are not missing completely at random (MCAR). Moreover, as the percentage of missing data becomes large, the deletion of all subjects with missing data is unnecessarily wasteful and inefficient.

Various statistical methods have been developed to estimate parameters of Cox proportional hazards model with missing covariates. Three common approaches will be reviewed here. The first one is the weighted estimating equations (WEE) method, the second is the likelihood-based method, and the third is the imputation method.

2.1 WEIGHTED ESTIMATING EQUATIONS METHOD

The weighted estimating equations (WEE) method is a special case of a method proposed earlier by Rubin, *et al.* [16]. With WEE, the contribution to the estimating equation from a complete observation is weighted by π, the inverse probability that the covariate is observed.

Lipsitz, Ibrahim and Zhao [5] developed WEE methods to estimate regression parameters $\boldsymbol{\beta}$ in generalized linear models with missing categorical and continuous covariates. The methods are quite general and can be applied to very large classes of models, including proportional hazards model and nonlinear models. Moreover, WEEs are almost identical to

maximum likelihood (ML) estimating methods, so that one could use an EM-type algorithm and extend Monte Carlo EM algorithm to a Monte Carlo WEE to solve weighted estimating equations. Let y_i be the outcome of ith subject, $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})'$ be the vector of p covariates that are always observed, z_i be a covariate that may be missing for some subjects, and r_i be the missing indicator which is equal to 1, if z_i is observed and 0, otherwise. The distribution of r_i given $(y_i, \mathbf{x}_i, z_i)$ follows Bernoulli distribution with probability $\pi_i = Pr(r_i = 1|y_i, \mathbf{x}_i, z_i)$ and is also called missing data mechanism. In Lipsitz, *et al.*, missing data mechanism is MAR or MCAR, so π_i does not depend on z_i. They proposed a logistic regression for the probability of being observed,

$$\pi_i = \pi_i(\mathbf{w}) = \frac{\exp(-\mathbf{w}'m_i)}{1 + \exp(-\mathbf{w}'m_i)}$$

where $\mathbf{w}$ is a vector of unknown parameters and m_i is some function of $(y_i, \mathbf{x}_i')$'s. Therefore, the score function of WEE corresponding to $\boldsymbol{\beta}$ is

$$\mathbf{u}^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{r_i}{\pi_i}\mathbf{u}_{1i}(\boldsymbol{\beta}; y_i, \mathbf{x}_i, z_i) + (1 - \frac{r_i}{\pi_i})E_{z_i|y_i,\mathbf{x}_i}[\mathbf{u}_{2i}(\boldsymbol{\beta}; y_i, \mathbf{x}_i, z_i)]$$

compared to that of ML,

$$\mathbf{u}^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} r_i\mathbf{u}_{1i}(\boldsymbol{\beta}; y_i, \mathbf{x}_i, z_i) + (1 - r_i)E_{z_i|y_i,\mathbf{x}_i}[\mathbf{u}_{2i}(\boldsymbol{\beta}; y_i, \mathbf{x}_i, z_i)] \ .$$

Using ML, the distribution of $(z_i|\mathbf{x}_i)$ and $(y_i, |\mathbf{x}_i, z_i)$ must be specified correctly in order to obtain a consistent estimate of $\boldsymbol{\beta}$. For the WEEs, the consistent estimate of $\boldsymbol{\beta}$ can be obtained as long as the model of π_i and distribution of $(y_i, |\mathbf{x}_i, z_i)$ are specified correctly. However, unlike ML, WEE requires a sufficient amount of missing data so that π_i can be estimated precisely.

Parzen, *et al.* [7] extended the previous methods to the situation when the distribution of missing covariates is misspecified as multivariate normal, but the WEE still generates consistent estimates of parameters of interest as long as the probability of being observed is correctly modeled. Moreover, the WEE can be used when the density of data is not normal and the missing covariates can be continuous and non-normal as well. Wang, *et al.* [24] used different estimates of the selection probabilities, π_i, to investigate the performance

of weighted estimators. In particular, they examined the properties of the estimates of the regression parameters when the selection probabilities are estimated by nonparametric smoothing, such as kernel and spline smoothers, so that misspecification of π_i can be avoided.

Wang and Chen [6] proposed the augmented inverse probability weighted estimator for Cox proportional hazards model with missing categorical covariates. If we let $T_i = \min(T_i^0, C_i)$ be the observed time and $\delta_i = I_{(T_i^0 \leq C_i)}$ be the failure indicator, which equals 1 if the event is observed and 0 otherwise. Also let X_i be a vector of covariates that may be missing in some subjects, Z_i is the vector of fully observed covariates, and r_i be the missing indicator that equals 1 if X_i is observed and zero otherwise. The AIPW estimator can be implemented by the EM-type algorithm, and it is doubly robust because it is consistent as long as either the selection probability model or the joint distribution of covariates is specified correctly.

However, the WEE method has some disadvantages compared with the maximum likelihood method, although it is more flexible in specifying the parametric density of covariates. The WEE method performs best when a vector of covariates that could be missing either has all elements observed or all elements unobserved, and the probability of being observed (π) must be correctly modeled. On the other hand, under MAR, the maximum likelihood method easily handles more than one missing covariate with any pattern of missingness, and mixed categorical and continuous covariates. The WEE method become more difficult because a model for the missing data mechanism must be specified and estimated, which is only easy to do when a set of covariates is always observed or always missing.

2.2 LIKELIHOOD-BASED METHOD

Previous work in the area of likelihood-based methods includes methods developed by Lin and Ying [8]. They were the first to introduce an estimator for the Cox proportional hazards model with missing covariates. The estimating function they proposed for the vector of regression parameters is the approximation partial-likelihood (APL) score function, which is an approximation to the partial likelihood score function with full covariate measurements.

When APL score function is equal to zero, estimators of parameters can be obtained by using the Newton-Raphson algorithm. Moreover, when the censoring is heavy, about 90%, the estimators are more efficient than the estimators based on completely observed data, but it is not the case when the censoring is less than 50%. In addition, it is biased under MAR.

Zhou and Pepe [25] proposed an estimated partial likelihood method to estimate relative risk parameters with a property that it is nonparametric with respect to the association between components of the missing and observed covariates. Although the method is semi-parametric and does not require strong assumptions, it has some limitations. For example, the covariates are required to be discrete, if the dimension of observed covariates is large, the method may lead to unstable estimates, and one needs to be careful while using some of assumptions.

Lipsitz and Ibrahim [26] developed a method for ignorable missing categorical covariates in parametric proportional hazards models. They recommended the use of estimating equations in situations where the missing data are missing at random and the pattern of missing data is monotone. However, the method is restrictive because the monotone missing data pattern seldom occurs in practice.

Chen and Little [27] proposed a nonparametric maximum likelihood (NPML) method to estimate regression parameters in a proportional hazard regression model with missing discrete covariates. The key feature of NPML is that the cumulative baseline hazard is a step function with jumps at follow-up times. Although the nonparametric "likelihood" is not standard, the NPML estimate has asymptotic properties similar with those of parametric likelihood estimates. Therefore, the EM algorithm and ECM algorithm are used to deal with the large number of parameters when the nonparametric likelihood is maximized. Their simulation study showed that, under MAR, NPML estimates of Cox proportional hazards model were efficient when covariates had missing elements. However, NPML has limitations:

1. NPML procedure works under MAR and missing covariates are all categorical or normally distributed ;
2. It limits missing-data pattern as monotonic; and
3. Model misspecification has not been investigated.

Lipsitz and Ibrahim [9] developed a likelihood-based approach to estimate regression coefficients in Cox proportional hazards model when there are missing categorical covariate data. If failure times T_i for subject i, $i = 1, 2, \ldots, n$, follows the Cox proportional hazards model, and the hazard function for T_i at time t is equation

$$\lambda(t) = \lambda_0(t) e^{\boldsymbol{\beta}' \mathbf{z}_i(t)}, \tag{2.1}$$

then the log-likelihood for parameter $\boldsymbol{\beta}$ is

$$l(\boldsymbol{\beta}; x_i, \delta_i, \mathbf{z}_i) = \delta_i log[\lambda_0(x_i)] + \delta_i(\boldsymbol{\beta}' \mathbf{z}_i) - e^{\boldsymbol{\beta}' \mathbf{z}_i} \Lambda_0(x_i)$$

where $X_i = \min(T_i, C_i)$ is the observed time, $\delta_i = I_{(T_i \leq C_i)}$ is the failure indicator, $\mathbf{z}_i$ is the vector of covariates, and

$$\Lambda_0(x_i) = \int_0^t \lambda_0(u) du$$

is the cumulative baseline hazard function. The score function of partial likelihood is given by $\mathbf{u}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ and

$$\bar{\mathbf{z}}(s, \boldsymbol{\beta}) = \frac{\sum_{j=1}^n \mathbf{z}_j Y_j(s) e^{\boldsymbol{\beta}' \mathbf{z}_j}}{\sum_{j=1}^n Y_j(s) e^{\boldsymbol{\beta}' \mathbf{z}_j}}$$

is a weighted average of the $\mathbf{z}_i$'s, and $dN_i = N_i(s) - N_i(s^-)$ is a binary random variable that equals one if subject i fails at time s and equals zero otherwise. When $\mathbf{u}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) = 0$, the root is the maximum partial likelihood estimate $\hat{\boldsymbol{\beta}}$. When there are missing values in covariates, $\mathbf{z}_i$ can be written as $\mathbf{z}_i = (\mathbf{z}_{mis,i}, \mathbf{z}_{obs,i})$, where $\mathbf{z}_{mis,i}$ is the missing components of $\mathbf{z}_i$ and $\mathbf{z}_{obs,i}$ is the observed components of $\mathbf{z}_i$. Therefore, under MAR, a consistent estimate of $\boldsymbol{\beta}$ can be obtained by letting the conditional expectation of score function given observed data equal zero and solving for $\hat{\boldsymbol{\beta}}$, that is

$$\mathbf{u}^*(\boldsymbol{\beta}) = E[\mathbf{u}(\boldsymbol{\beta}) | \text{observed data}] = E[\mathbf{u}(\boldsymbol{\beta}) | (\mathbf{z}_{obs,1}, x_1, \delta_1), \ldots, (\mathbf{z}_{obs,n}, x_n, \delta_n)] = 0 \ . \tag{2.2}$$

Note that, the expectation in (2.2) is taken with respect to the conditional distribution of the missing data given the observed data $(x_i, \delta_i, \mathbf{z}_{obs,i})$ for subject i. Because the missing covariates are discrete, the equation (2.2) can be written as

$$\mathbf{u}^*(\boldsymbol{\beta}) = \sum_{\mathbf{z}_{mis,1(j)}}^{n1} \cdots \sum_{\mathbf{z}_{mis,n(j)}}^{nn} p_{1j} \ldots p_{nj} \left[\sum_{i=1}^n \int_0^\infty \{\mathbf{z}_i - \bar{\mathbf{z}}(s, \boldsymbol{\beta})\} dN_i(s) \right]$$

where for i subject, p_{ij} is the conditional distribution of a particular missing data pattern, indexed by j, given observed data and also can be viewed as the posterior probabilities of the missing covariates. Thus,

$$\begin{aligned} p_{ij} &= Pr[\mathbf{z}_{mis,i} = \mathbf{z}_{mis,i}(j)|(\mathbf{z}_{obs,i}, x_i, \delta_i, \boldsymbol{\beta})] \\ &= \frac{p(x_i, \delta_i|\mathbf{z}_i, \lambda, \boldsymbol{\beta})p(\mathbf{z}_i|\boldsymbol{\alpha})}{\sum_{\mathbf{z}_{mis,i}} p(x_i, \delta_i|\mathbf{z}_i, \lambda, \boldsymbol{\beta})p(\mathbf{z}_i|\boldsymbol{\alpha})} \\ &= \frac{p(x_i, \delta_i|\mathbf{z}_{mis,i}(j), \mathbf{z}_{obs,i}, \lambda, \boldsymbol{\beta})p(\mathbf{z}_{mis,i}(j), \mathbf{z}_{obs,i}|\boldsymbol{\alpha})}{\sum_{\mathbf{z}_{mis,i}} p(x_i, \delta_i|\mathbf{z}_i, \lambda, \boldsymbol{\beta})p(\mathbf{z}_i|\boldsymbol{\alpha})} \end{aligned} \tag{2.3}$$

where $j(= 1, \ldots, n_i)$ is the n_i distinct covariate patterns that $\mathbf{z}_{mis,i}$ could be in given $(\mathbf{z}_{obs,i}, x_i, \delta_i)$, $\mathbf{z}_{mis,i}(j)$ is the jth possible missing data pattern for subject i, and $\boldsymbol{\alpha}$ is the vector of parameters of the distribution of covariates. Because there are missing values in covariates, the distribution of covariates is no longer ancillary when estimating $\boldsymbol{\beta}$, the distribution of $\mathbf{z}_{mis,i}$ needs to be specified. Therefore, the score function of the joint distribution of covariates is given by

$$\mathbf{u}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{\partial log p(\mathbf{z}_i|\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \ .$$

In addition, a nonparametric estimate of the cumulative baseline hazard function $\Lambda_0(t)$ needs to be specified as well. Therefore, an EM-type algorithm that is similar to the EM algorithm was proposed to solve the estimating equations $\mathbf{u}^*(\hat{\boldsymbol{\theta}}) = 0$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \Lambda_0(t), \boldsymbol{\alpha})$ and the score function has the form of

$$\mathbf{u}^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) = \sum_{\mathbf{z}_{mis,1(j)}}^{n1} \cdots \sum_{\mathbf{z}_{mis,n(j)}}^{n_n} p_{1j}^{(1)} \cdots p_{nj}^{(m)} \begin{bmatrix} \mathbf{u}_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(m)}) \\ \mathbf{u}_{\Lambda}(\boldsymbol{\Lambda_0(t)}|\boldsymbol{\theta}^{(\mathbf{m})}) \\ \mathbf{u}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}|\boldsymbol{\theta}^{(m)}) \end{bmatrix}$$

where $p_{ij}^{(m)} = p_{ij}(\boldsymbol{\theta}^{(m)})$ is the posterior distribution in (2.3) estimated at the m step of given value $\boldsymbol{\theta}$. Furthermore, a Monte Carlo approximation is developed to reduce the computational burden of obtaining estimates. In the Monte Carlo algorithm, given the estimate $\boldsymbol{\theta}^{(m)}$, L values of $\mathbf{z}_{mis,i}$ are drawn from the conditional distribution of $\mathbf{z}_{mis,i}$ given the observed data $(\mathbf{z}_{obs,i}, x_i, \delta_i, \boldsymbol{\theta}^{(m)})$, where the lth draw is denoted by $\mathbf{z}_{mis}^{l(m)}$. Thus, in iterations of the

Monte Carlo algorithm, score function $\mathbf{u}^{**}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) = 0$ is solved to obtain the estimates of parameters. Where

$$\mathbf{u}^{**}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) = \frac{1}{L}\sum_{l=1}^{L}\mathbf{u}(\boldsymbol{\theta}, \mathbf{z}_{mis}^{l(m)})$$

In this paper, the authors focused on general patterns of missingness, compared with the method proposed by Zhou and Pepe [25] and Chen and Little [27], which can only be used with monotone missingness.

Ibrahim, *et al.* [28] extended the method to general parametric regression models with missing continuous covariates, allowed any pattern of missing data and assumed that the missing data mechanism is non-ignorable.

Herring and Ibrahim [10] presented the Monte Carlo EM methods to estimate parameters of interest in Cox proportional hazards model when missing covariates are categorical, continuous and mixed. The missing data mechanism is missing at random. This paper is the extention of Lipsitz and Ibrahim [9]. When missing covariates are continuous, they proposed Gibbs sampler (Gelfand and Smith [52]) along with the adaptive rejection algorithm of Gilks and Wild [30] to take a sample from posterior distributions of missing covariates, then fill in the missing data, and use Monte Carlo method to simplify the score function with inverse of the number of missing data patterns as weights. The technique to mixed types of covariates is similar with the one mentioned in Ibrahim, *et al.* [28]. That is, for the ith subject, $\mathbf{z}_{mis,i}$ can be defined as $\mathbf{z}_{mis,i} = (\mathbf{z}_{mis,i}^{d}, \mathbf{z}_{mis,i}^{c})$, where $\mathbf{z}_{mis,i}^{d}$ denotes the categorical covariates in $\mathbf{z}_{mis,i}$ and $\mathbf{z}_{mis,i}^{c}$ denoted the continuous covariates in $\mathbf{z}_{mis,i}$. Note that other notation follows that of Lipsitz and Ibrahim [9]. For the ith subject, the expectation of the score contribution is summing over the categorical covariates

$$\begin{aligned}\tilde{\mathbf{u}}_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(m)})_i &= \sum_{\mathbf{z}_{mis,i}^{d}}\int\int_{0}^{\infty}[\{\mathbf{z}_i - \bar{\mathbf{z}}(u,\boldsymbol{\beta})dN_i(u)\}\times p(\mathbf{z}_{mis,i}^{d}(j)|\mathbf{z}_{mis,i}^{c}, \mathbf{z}_{obs,i}, x_i, \delta_i, \boldsymbol{\theta}^{(m)}) \\ &\quad \times p(\mathbf{z}_{mis,i}^{c}|\mathbf{z}_{obs,i}, x_i, \delta_i, \boldsymbol{\theta}^{(m)})d\mathbf{z}_{mis,i}^{c}\end{aligned}$$

where

$$\bar{\mathbf{z}}_w(u,\boldsymbol{\beta}) = \frac{\sum_{i=1}^{n}\{(1/n_i)\sum_{k=1}^{n_i}\mathbf{z}_{i,k}^{*}Y_i(u)\exp(\boldsymbol{\beta}'\mathbf{z}_{i,k}^{*})\}}{\sum_{i=1}^{n}\{(1/n_i)\sum_{k=1}^{n_i}Y_i(u)\exp(\boldsymbol{\beta}'\mathbf{z}_{i,k}^{*})\}}\ .$$

In addition, a sample $(s^c_{i,1}, \ldots, s^c_{i,n_i})$ of size n_i is drawn from $p(\mathbf{z}^c_{mis,i}|\mathbf{z}_{obs,i}, x_i, \delta_i, \boldsymbol{\theta}^{(m)})$ by using the Gibbs sampler along with the adaptive rejection algorithm. Then the Monte Carlo estimate is calculated from the expectation of the score function, which is

$$\tilde{\mathbf{u}}_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\boldsymbol{\theta}^{(m)})_i = \frac{1}{n_i} \sum_{\mathbf{z}^d_{mis,i(j)}} \sum_{k=1}^{n_i} \int p^{(m)}_{ijk}[\mathbf{z}^*_{i,k} - \bar{\mathbf{z}}_w(\boldsymbol{\beta}, u)] dN_i(u)$$

where $\mathbf{z}^*_{i,k} = (\mathbf{z}_{obs,i}, s^c_{i,k})$ and the weights are given by

$$p^{(m)}_{ijk} = p(\mathbf{x}^d_{mis,i}(j)|s^c_{i,k}, \mathbf{z}_{obs,i}, x_i, \delta_i, \boldsymbol{\theta}^{(m)}) \ .$$

In later simulations, the authors compared the estimators of their methods with those of previous methods proposed by Lin and Ying [8], Paik and Tsai [11] and Chen and Little [27]. The results are favor to the estimators of their methods from both efficiency and bias correction points of view. The proposed methodology by Herring and Ibrahim [10] is very general, which allows any missing data pattern and works under missing at random, but the computation is quite intensive.

2.3 IMPUTATION METHOD

Paik and Tsai [11] proposed a single imputation method to estimate parameters of Cox proportional hazards model with missing covariates under MAR. Their suggestion was to impute the conditional expectation of the statistic involving missing covariates given observed information. They introduced a consistent estimator that was a solution of an approximated partial likelihood equation, which was obtained by replacing the missing covariates with the observed covariates from the same risk set. Therefore, the information due to missing covariates can be regained from observed covariates with appropriate conditions. However, it becames really messy and imprecise when one tries to choose imputations from observed variables for missing terms, especially, when the missing covariates are continuous, although smoothing technique was employed.

Multiple imputation [20] is a general method developed from a Bayesian perspective for handling missing-data problems. The mechanism is to create multiple completed data sets

by imputing plausible values for the missing data, analyze each filled-in data set as though it were the complete data, then combine inferences into a single overall inference. Paik [21] used a multiple imputation method called Approximate Bayesian Bootstrap (ABB) to deal with a simple scenario when only one covariate was missing. The proposed multiple imputation estimates had a simpler variance calculation than the previous imputation-based estimates, but with a small loss in efficiency. The method discussed in this article assumed that all components of the covariates were categorical, when any component of covariates was continuous, smoothing technique as suggested by Paik and Tsai [11] may be used. However, the smoothing technique would introduce a bias. Buuren, et al. [12] applied multiple imputation method to impute missing values of blood pressure covariates in survival analysis. For the univariate case, covariates were modeled by a linear regression model. Random draws from posterior distributions of regression parameters were computed and filled-in the missing values. The procedure was repeated m times. For multivariate case, instead of assuming a conditional distribution for each missing variable, the authors did not explicitly assume a specific distribution as in Schafer [18], but assumed that the distribution exists. Draws from conditional distributions were generated by using Gibbs sampling. Let Y_1, Y_2, Z, V, and U be covariates that may be missing in some subjects. First, a random draw from marginal distributions of observed variables is filled in each missing value. Then Y_1 is imputed by the procedure conditional on all other data (observed and imputed combined), then Y_2 conditional on all other data (using most recent imputed Y_1), and so on, until all missing variables in Y_1, Y_2, Z, V, and U have been imputed. Subsequent iteration of the process. However, the method proposed to generate the imputations has some limitations that could affect relative effects of predictors. Cho and Schenker [22] estimated parameters in the accelerated failure time (AFT) model via Gibbs sampling when there were missing values in covariates. They used a general location model and allowed different covariance structures for continuous covariates across categories of discrete covariates. Gibbs sampling was applied to create multiple imputations for missing values. Moreover, they suggested that the proposed methods can be used on Cox proportional hazards model with missing covariates. However, they were concerned that the imputation and analysis strategies would be an issue for implementing other models. Schafer's book [18] presents iterative algorithms

for simulating multiple imputations of missing values in missing datasets under multivariate models. Categorical covariates are assumed to be from a multinomial distribution, and continuous covariates are assumed to be from a multivariate normal distribution. The proposed approach allows one to analyze the data by any technique that would be appropriate if the data were complete. In addition, the book introduces the general location model (GLLM) to deal with missing categorical covariates, continuous covariates or mixed types of covariates. Moreover, the methods presented in the book were based on two distinct concepts: one is likelihood-based inference with missing data, in particular, the EM algorithm; the other is the technique of Markov Chain Monte Carlo, particularly, data augmentation. The EM algorithm is used to compute maximum-likelihood estimates (MLE) for parameters of the general location model for an incomplete dataset, and the MLEs would be used as starting values of the parameters for next iteration. Data augmentation, which has I-step and P-step, is used to generate posterior draws of the parameters of GLLM given missing mixed data, and imputations of missing covariates are generated by the I-step of data augmentation. At the I-step, missing data are randomly imputed by random draws from predictive distributions given observed data and current parameter values; and at the P-step, a new parameter value is drawn from its posterior distribution given the completed data. Multiple imputations are obtained by repeating data augmentation m times, so m sets of complete data are created. In addition, when the sample size, n, is relatively small, and the total number of categories is relatively large, restricted GLLM can be used instead of unrestricted GLLM. The algorithms proposed in the book are computationally intensive, but they have been implemented by the author for general use as functions in the statistical package, and can be downloaded and installed in statistical software such as S-Plus and R. The convenient computation makes Schafer's approaches more attractive and practical than other approaches reviewed here.

In summary, the weighted estimating equations method, the likelihood-based method and the multiple imputation method have been investigated by many researchers. The weighted estimating equations method has more practical disadvantages than the other two methods. In particular, it has a good estimation only when a set of covariates is always observed or always missing, which limits applications of this method. On the other hand, the likelihood-based method has been developed as a general method, but it is theoretically complicated.

Moreover, the computation is quite intensive. In contrast, the multiple imputation method has a relatively simple concept, and it is a general method as well. Furthermore, some of proposed multiple imputation methods are written as functions of statistical packages by their authors, and can be used easily in statistical software, which makes the method more practical, easy and widely used than other two methods. In the next chapter, we will compare and evaluate two multiple imputation algorithms, both implemented in R, namely, Schafer's MIX [18] (estimation/multiple imputation for Mixed categorical and continuous data) and Buuren and Oudshoorn's MICE ($V1.0$) [36] (Multiple Imputation by Chain Equations). They implement multiple imputation for mixed types of missing covariates, that is, missing values can be present in both continuous and categorical covariates. Although Horton and Lipsitz [35] reviewed some multiple imputation packages, they did not assess the performance of MICE under a large amount of missingness and nonmonotone missing data patterns for incomplete covariates in survival analysis, and Schafer's software was not evaluated and compared with the other packages.

3.0 INTRODUCTION TO MULTIPLE IMPUTATION METHOD FOR MISSING COVARIATES

Multiple imputation [20] is a general method developed from a Bayesian perspective for handling missing data problems. It is a three-step procedure. The first step is to generate m sets of plausible values for missing observations from an appropriate imputation model and fill them back into the original data, thus having m complete datasets. Second, each complete dataset can be analyzed using specific methods or models. Finally, in the third step, the results of the m analyses are combined in a simple way that accounts for the uncertainty regarding the imputation process.

3.1 NOTATION AND CONCEPTS FOR COMPLETE DATA

Cox proportional hazards models were proposed by Cox [1], and allow one to quantify the relationship between the time to event and a set of covariates. It is a semiparametric regression model and assumes that hazard function for the failure time T, conditional on covariates $\mathbf{z}$, is the form

$$\lim_{h\to 0} \frac{P(t \leq T < t+h | T \geq t, \mathbf{z})}{h} = \lambda(t|\mathbf{z}) = \lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{z}) \tag{3.1}$$

where $\mathbf{z}$ is a p vector of covariates, $\lambda_0(t)$ is an unspecified baseline hazard function, and $\boldsymbol{\beta}$ is a parameter vector with dimension $p \times 1$. The covariate vector, $\mathbf{z}$, may be categorical or continuous.

The density of failure time T_i can by written as

$$\begin{aligned} p(t_i|\mathbf{z}_i, \boldsymbol{\beta}) &= \lambda(t_i|\mathbf{z}_i, \boldsymbol{\beta}) S(t_i|\mathbf{z}_i, \boldsymbol{\beta}) \\ &= \lambda_0(t_i) \exp(\boldsymbol{\beta}'\mathbf{z}_i) \exp(-\exp((\boldsymbol{\beta}'\mathbf{z}_i)\Lambda_0(t_i)) \end{aligned}$$

where

$$S(t_i|\mathbf{z}_i, \boldsymbol{\beta}) = pr(T_i > t_i|\mathbf{z}_i, \boldsymbol{\beta}) = \exp(-\exp((\boldsymbol{\beta}'\mathbf{z}_i)\Lambda_0(t_i))$$

is the survivor function for T_i, and

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du$$

is the cumulative baseline hazard function.

Let $X(t) = \min(T, U^c)$ is the observed event time, where T is the failure time and U^c is the censoring time, and $\delta = I_{(T \leq X(t))}$ is the failure indicator that is equal to one if the observed event is failure and zero otherwise. The full data are expressed as $(x_i(t), \delta_i, \mathbf{z}_i)$, $i = 1, \ldots, n$. Therefore, the probability distribution for $(x_i(t), \delta_i|\mathbf{z}_i)$ is

$$\begin{aligned} p(x_i(t), \delta_i|\mathbf{z}_i, \boldsymbol{\beta}) &\propto p(x_i(t)|\mathbf{z}_i, \boldsymbol{\beta})^{\delta_i} S(x_i(t)|\mathbf{z}_i, \boldsymbol{\beta})^{1-\delta_i} \\ &= \lambda(x_i(t)|\mathbf{z}_i, \boldsymbol{\beta})^{\delta_i} S(x_i(t)|\mathbf{z}_i, \boldsymbol{\beta})^{1-\delta_i} \\ &= [\lambda_0(x_i(t)) \exp(\boldsymbol{\beta}'\mathbf{z}_i)]^{\delta_i} \exp(-\exp((\boldsymbol{\beta}'\mathbf{z}_i)\Lambda_0(x_i(t))) \end{aligned}$$

where the censoring is noninformative in that, given $\mathbf{z}_i$, the event and censoring time for the ith subject are independent. Thus the log-likelihood for subject i is given by

$$\begin{aligned} l(\boldsymbol{\beta}|x_i(t), \delta_i, \mathbf{z}_i) &= \delta_i \log[p(x_i(t)|\mathbf{z}_i, \boldsymbol{\beta})] + (1 - \delta_i) \log[S(x_i(t)|\mathbf{z}_i, \boldsymbol{\beta})] \\ &= \delta_i(\log \lambda_0(x_i(t)) + \boldsymbol{\beta}'\mathbf{z}_i) - \exp(\boldsymbol{\beta}'\mathbf{z}_i)\Lambda_0(x_i(t)) \ . \end{aligned}$$

Generally, in order to avoid the estimation of unspecified $\lambda_0(t)$ and $\Lambda_0(t)$ in the log-likelihood, the partial likelihood based on the hazard function as defined by (3.1) can be used. If there are no ties between the event times, the partial likelihood is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}'\mathbf{z}_i)}{\sum_{j=1}^{n} I_{\{x_j(t) \geq x_i(t)\}} \exp(\boldsymbol{\beta}'\mathbf{z}_j)} \right]^{\delta_i} ,$$

and the estimate of $\boldsymbol{\beta}$ can be obtained by setting partial likelihood score function to zero. The information matrix is the negative of the matrix of second derivatives of the log likelihood and is given by $I^{-1}(\boldsymbol{\beta})$. Under assumptions of noinformative and independent censoring, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})$ is asymptotically distributed as a multivariate normal distribution with mean 0 and covariance matrix $nI^{-1}(\hat{\boldsymbol{\beta}})$.

Alternatively, using counting process notation, the processes $\{N_i(t), Y_i(t) : t \geq 0\}$ represents the information contained in $(x_i(t), \delta_i)$, where $N_i(t) = I_{\{T_i \leq t, \delta_i = 1\}}$ and $Y_i(t) = I_{\{x_i(t) \geq t\}}$, the score function can be written as

$$\mathbf{u}_\beta(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\infty \{\mathbf{z}_i - \bar{\mathbf{z}}(\boldsymbol{\beta}, u)\} dN_i(u) \tag{3.2}$$

where

$$\begin{aligned} \bar{\mathbf{z}}(\boldsymbol{\beta}, u) &= \frac{\sum_{j=1}^{n} \mathbf{z}_j Y_j(u) e^{\boldsymbol{\beta}' \mathbf{z}_j}}{\sum_{j=1}^{n} Y_j(u) e^{\boldsymbol{\beta}' \mathbf{z}_j}} \\ &= \frac{S^{(1)}(\boldsymbol{\beta}, u)}{S^{(0)}(\boldsymbol{\beta}, u)} \ . \end{aligned}$$

When $\mathbf{u}_\beta(\hat{\boldsymbol{\beta}}) = 0$, the solution is the maximum partial likelihood estimate $\hat{\boldsymbol{\beta}}$.

3.2 MODEL ASSUMPTIONS FOR MISSING COVARIATE ESTIMATION

3.2.1 The complete data model for covariates

Let $\mathbf{Z}$ denote the $n \times p$ matrix of complete covariate data, which is not fully observed, $\mathbf{z}_i$ denote the i row of $\mathbf{Z}$, $i = 1, \ldots, n$, and the rows are independent, identically distributed (iid) draws from some multivariate probability distribution, where n represents the number of subjects, and p represents the number of covariates of each subject. The probability density of complete covariate data can be given by

$$P(\mathbf{Z}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{z}_i|\boldsymbol{\theta}) \ ,$$

where f is the density for a single row, and $\boldsymbol{\theta}$ is a vector of unknown parameters. Moreover, distribution of f will be assumed to be from one of three classes:

1. the multivariate normal distribution for continuous covariates;
2. the multinomial model for categorical covariates; and
3. a class of models for mixed normal and categorical covariates (Little and Schluchter [32]).

3.2.2 Ignorability

3.2.2.1 Missing at random Let $\mathbf{Z}_{obs}$ denote observed part of $\mathbf{Z}$, and $\mathbf{Z}_{mis}$ denote missing part of $\mathbf{Z}$, so that $\mathbf{Z} = (\mathbf{Z}_{obs}, \mathbf{Z}_{mis})$. We assume that the missing data mechanism is MAR throughout this dissertation. MAR was defined by Rubin [34], and it means that the probability that an observation is missing may depend on the components of observed data ($\mathbf{Z}_{obs}$), but not on the components that are missing ($\mathbf{Z}_{mis}$). More formally, MAR in terms of a probability model for the missingness can be written as

$$f(\mathbf{R}|\mathbf{Z}, \boldsymbol{\tau}) = f(\mathbf{R}|\mathbf{Z}_{obs}, \boldsymbol{\tau})$$

for all $\mathbf{Z}_{mis}$ and $\boldsymbol{\tau}$, where $\mathbf{R} = (R_1, \ldots, R_n) = (r_{ij})$, $i = 1, \ldots, n$ and $j = 1, \ldots, p$, be an $n \times p$ matrix of missing-data indicators, and $r_{ij} = 1$ if the elements of $\mathbf{Z}$ are observed and $r_{ij} = 0$ if the elements of $\mathbf{Z}$ are missing. $\boldsymbol{\tau}$ represents the vector of unknown parameters.

3.2.2.2 Distinctness of parameters The parameter, $\boldsymbol{\theta}$, of the data model and the $\boldsymbol{\tau}$ of the missing-data mechanism are assumed *distinct*, which is defined as that the joint parameter space of $(\boldsymbol{\theta}, \boldsymbol{\tau})$ is the product of the parameter space of $\boldsymbol{\theta}$ and the parameter space of $\boldsymbol{\tau}$. Therefore, if both MAR and distinctness hold, then the missing-data mechanism is said to be *ignorable* (Little and Rubin [4]).

3.2.3 Likelihood-based inference with missing data

Rubin [34] and Little and Rubin [4] point out that we do not need to consider the model for $\mathbf{R}$ nor the nuisance parameters $\boldsymbol{\tau}$ under ignorability when making likelihood-based or Baysian inference about $\boldsymbol{\theta}$.

The distribution of the observed data is obtained by integrating $\mathbf{Z}_{mis}$ out of the joint distribution of $\mathbf{Z} = (\mathbf{Z}_{obs}, \mathbf{Z}_{mis})$ and $\mathbf{R}$, that is

$$f(\mathbf{Z}_{obs}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\tau}) = \int f(\mathbf{Z}_{obs}, \mathbf{Z}_{mis}|\boldsymbol{\theta}) f(\mathbf{R}|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}, \boldsymbol{\tau}) d\mathbf{Z}_{mis} \ . \tag{3.3}$$

Under the MAR assumption, equation (3.3) becomes

$$\begin{aligned} f(\mathbf{Z}_{obs}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\tau}) &= f(\mathbf{R}|\mathbf{Z}_{obs}, \boldsymbol{\tau}) \int f(\mathbf{Z}_{obs}, \mathbf{Z}_{mis}|\boldsymbol{\theta}) d\mathbf{Z}_{mis} \\ &= f(\mathbf{R}|\mathbf{Z}_{obs}, \boldsymbol{\tau}) f(\mathbf{Z}_{obs}|\boldsymbol{\theta}) \ . \end{aligned} \tag{3.4}$$

When the missing-data mechanism is ignorable, the likelihood-based inference for $\boldsymbol{\theta}$ from $L(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{Z}_{obs}, \mathbf{R})$ will be the same as likelihood-based inference for $\boldsymbol{\theta}$ from

$$L_{ign}(\boldsymbol{\theta}|\mathbf{Z}_{obs}) \propto f(\mathbf{Z}_{obs}|\boldsymbol{\theta}) \ .$$

That is, the likelihood of $\boldsymbol{\theta}$ based on data $\mathbf{Z}_{obs}$ *ignoring the missing-data mechanism* is any function of $\boldsymbol{\theta}$ proportional to $f(\mathbf{Z}_{obs}|\boldsymbol{\theta})$.

In addition, Bayes inference for $\mathbf{Z}$ and $\mathbf{R}$ is given by combining the likelihood

$$L(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{Z}_{obs}, \mathbf{R}) \propto f(\mathbf{Z}_{obs}, \mathbf{R}|\boldsymbol{\theta}, \boldsymbol{\tau})$$

with a prior distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$:

$$P(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{Z}_{obs}, \mathbf{R}) \propto p(\boldsymbol{\theta}, \boldsymbol{\tau}) \times L(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{Z}_{obs}, \mathbf{R}) \ . \tag{3.5}$$

Under MAR and distinctness, the prior of $\boldsymbol{\theta}$ is independent of that of $\boldsymbol{\tau}$, $p(\boldsymbol{\theta}, \boldsymbol{\tau}) = p(\boldsymbol{\theta})p(\boldsymbol{\tau})$, then equation (3.5) can be written as

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{Z}_{obs}, \mathbf{R}) &\propto [p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\mathbf{Z}_{obs})][p(\boldsymbol{\tau})L(\boldsymbol{\tau}|\mathbf{Z}_{obs}, \mathbf{R}] \\ &\propto P(\boldsymbol{\theta}|\mathbf{Z}_{obs}) P(\boldsymbol{\tau}|\mathbf{Z}_{obs}, \mathbf{R}) \ . \end{aligned}$$

Thus the inference about $\boldsymbol{\theta}$ can be given by the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Z}_{obs})$ ignoring the missing-data mechanism.

3.2.4 EM Algorithm

The EM algorithm is an iterative algorithm for maximum likelihood (ML) estimation in missing-data problems. It consists of an E-step and an M-step. The E-step finds the conditional expectation of the missing data given the observed data and current estimated parameters, and then substitutes these expectations for the missing data; in M-step, ML estimation of $\boldsymbol{\theta}$ given the observed and the filled-in data is performed. The two steps are iterated until the estimates converge. The key idea of EM is that "missing data" are not $\mathbf{Z}_{mis}$ but the functions of $\mathbf{Z}_{mis}$ in the complete-data loglikelihood.

Specifically, the E-step of EM is to find the expected complete-data loglikelihood if $\boldsymbol{\theta} = \boldsymbol{\theta}^t$, where $\boldsymbol{\theta}^t$ is the current or t^{th} estimate of the parameters:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \int l(\boldsymbol{\theta}|\mathbf{Z}) f(\mathbf{Z}_{mis}|\mathbf{Z}_{obs}, \boldsymbol{\theta} = \boldsymbol{\theta}^t) d\mathbf{Z}_{mis} \ .$$

Note that, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$ is the expectation of loglikelihood $l(\boldsymbol{\theta}|\mathbf{Z})$ with respect to the conditional predictive distribution of $\mathbf{Z}_{mis}$.

The M-step of EM maximizes this expected complete-data loglikelihood to determine $\boldsymbol{\theta}^{t+1}$, so that

$$Q(\boldsymbol{\theta}^{t+1}|\boldsymbol{\theta}^t) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) \ ,$$

for all $\boldsymbol{\theta}$.

3.2.5 Data Augmentation

Data augmentation [33] is an iterative method of simulating the posterior distribution of $\boldsymbol{\theta}$ that combines features of the EM algorithm and multiple imputation. It has two steps that can be viewed as a small-sample refinement of the EM algorithm : the imputation (or I) step corresponding to the E-step and the posterior (or P) step corresponding to the M-step. The algorithm starts with an initial draw $\boldsymbol{\theta}^0$ from an approximation to the posterior distribution of $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}^t$ be a vector of $\boldsymbol{\theta}$ at iteration t. Then the algorithm is written as

I-step: Draw $\mathbf{Z}_{mis}^{t+1}$ from the distribution $P(\mathbf{Z}_{mis}|\mathbf{Z}_{obs}, \boldsymbol{\theta}^t)$;

P-step: Draw $\boldsymbol{\theta}^{t+1}$ from the distribution $P(\boldsymbol{\theta}|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}^{t+1})$.

At each step, missing data are randomly imputed from their posterior predictive distribution given the current parameter and observed data, and the updated parameter is drawn from its posterior distribution given the complete data.

The motivation of the procedure is that the distributions in these two steps are much easier to draw from than either of the posterior distributions $P(\mathbf{Z}_{mis}|\mathbf{Z}_{obs})$ and $P(\boldsymbol{\theta}|\mathbf{Z}_{obs})$, or the joint posterior distribution $P(\boldsymbol{\theta}, \mathbf{Z}_{mis}|\mathbf{Z}_{obs})$.

3.3 INTRODUCTION TO MIX

The multiple imputation algorithm for MIX is based on the general location model under MAR, and multiple imputations are generated by EM algorithm and Data Augmentation.

3.3.1 The general location model

3.3.1.1 Definition For n subjects, covariate $\mathbf{Z} = (\mathbf{Q}, \mathbf{Y})$ is an $n \times (p + w)$ matrix. Let $\mathbf{Q} = (\mathbf{Q_1}, \mathbf{Q_2}, \ldots, \mathbf{Q_p})$ denote a set of categorical covariates and $\mathbf{Y} = (\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_w})$ denote a set of continuous ones, and $\mathbf{q}_i$ and $\mathbf{y}_i$ denote vectors of the values of $\mathbf{Q}$ and $\mathbf{Y}$, respectively, for subject i, $i = 1, \ldots, n$. The cross-classification on the components of $\mathbf{Q}$ summarizes as a p-dimensional contingency table with $C = \prod_{j=1}^{p} c_j$ cells, where $I_j = 1, \ldots, c_j$ be the possible categories that Q_j takes. The cells of the contingency table can be arranged in a linear order indexed by $c = 1, \ldots, C$. Cell counts of subjects denote as $\mathbf{x} = \{x_c : c = 1, \ldots, C\}$, and $\mathbf{x}$ will be viewed as either a multidimensional array or a vector depending on the context. Let $\mathbf{U}$ be an $n \times C$ matrix with rows $\mathbf{u}_i^T$, $i = 1, \ldots, n$ and "T" represents the transpose of a vector or matrix . Here $\mathbf{u}_i^T = \mathbf{E_c}$ is a $1 \times C$ indicator-vector with 1 if subject i belongs to cell c of the contingency table, and 0's elsewhere. Thus, each row of $\mathbf{U}$ is not observed unless all categorical variables are observed, and $\mathbf{U^T U}$ is a $C \times C$ matrix with $\mathbf{x} = \{x_1, \ldots, x_C\}$ in the diagonals.

The general location model is defined for mixtures of continuous and categorical variables

by Olkin and Tate [38]. It is given by

$$(\mathbf{x}|\boldsymbol{\pi}) \sim M(n, \boldsymbol{\pi}), \tag{3.6}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_C)^T$, and

$$(\mathbf{y}_i|\mathbf{u}_i = \mathbf{E}_c, \boldsymbol{\mu}_c, \boldsymbol{\Omega}) \sim MN(\boldsymbol{\mu}_c, \boldsymbol{\Omega}). \tag{3.7}$$

Equation (3.6) is the marginal distribution of $\mathbf{Q}$ which is a multinomial distribution represented by cell counts x_c given cell probabilities $\pi_c = Pr(\mathbf{u}_i = \mathbf{E}_c)$ with $\sum \pi_c = 1$, for $i = 1, \ldots, n$ and $c = 1, \ldots, C$; equation (3.7) is the conditional distribution of $\mathbf{Y}$ given $\mathbf{Q}$ which is a multivariate normal distribution with $C \times w$ mean matrix $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C)^T$, where $\boldsymbol{\mu}_c$ is a w-vector of means corresponding to cell c, and $\boldsymbol{\Omega}$ is $w \times w$ covariance matrix . A common covariance structure and no structure zero are assumed for all cells. In addition, the model for $\mathbf{Y}$ given $\mathbf{Q}$ can be regarded as a multivariate regression,

$$\mathbf{Y} = \mathbf{U}\boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

where a $\boldsymbol{\varepsilon}$ is a $n \times w$ matrix of errors and the rows of the matrix are distributed independently as $N(\mathbf{0}, \boldsymbol{\Omega})$.

Therefore, parameters of the general location model are

$$\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Omega})$$

3.3.1.2 The likelihood function The likelihood function of the general location model is written as

$$L(\boldsymbol{\theta}|\mathbf{Y}) \propto \mathbf{L}(\boldsymbol{\pi}|\mathbf{Q})\mathbf{L}(\boldsymbol{\mu}, \boldsymbol{\Omega}|\mathbf{Q}, \mathbf{Y}),$$

it is the product of multinomial and normal likelihoods, which are

$$L(\boldsymbol{\pi}|\mathbf{Q}) \propto \prod_{\mathbf{c=1}}^{\mathbf{C}} \pi_{\mathbf{c}}^{\mathbf{x_c}}$$

and

$$L(\boldsymbol{\mu}, \boldsymbol{\Omega}|\mathbf{Q}, \mathbf{Y}) \propto |\boldsymbol{\Omega}|^{-\frac{\mathbf{n}}{\mathbf{2}}} \exp\left\{-\frac{\mathbf{1}}{\mathbf{2}} \sum_{\mathbf{c=1}}^{\mathbf{C}} \sum_{\mathbf{i} \in \mathbf{F_c}} (\mathbf{y_i} - \boldsymbol{\mu}_\mathbf{c})^{\mathbf{T}} \boldsymbol{\Omega}^{-\mathbf{1}} (\mathbf{y_i} - \boldsymbol{\mu}_\mathbf{c})\right\}$$

where F_c is the set of all subjects fall into the cell c. And

$$L(\boldsymbol{\mu}, \boldsymbol{\Omega}|\mathbf{Q}, \mathbf{Y}) \propto |\boldsymbol{\Omega}|^{-\frac{\mathbf{n}}{\mathbf{2}}} \exp\left\{-\frac{\mathbf{1}}{\mathbf{2}} \mathrm{tr} \boldsymbol{\Omega}^{-1} \mathbf{Y}^{\mathbf{T}} \mathbf{Y} + \mathrm{tr} \boldsymbol{\Omega}^{-1} \boldsymbol{\mu}^{\mathrm{T}} \mathbf{U}^{\mathbf{T}} \mathbf{Y} - \frac{\mathbf{1}}{\mathbf{2}} \mathrm{tr} \boldsymbol{\Omega}^{-1} \boldsymbol{\mu}^{\mathrm{T}} \mathbf{U}^{\mathbf{T}} \mathbf{U} \boldsymbol{\mu}\right\},$$

so the likelihood is linear in the sufficient statistics $\mathbf{S}_1 = \mathbf{Y}^{\mathbf{T}}\mathbf{Y}$, $\mathbf{S}_2 = \mathbf{U}^{\mathbf{T}}\mathbf{Y}$, and $\mathbf{S}_3 = \mathbf{U}^{\mathbf{T}}\mathbf{U}$.

3.3.1.3 Prior Distributions

It is also convenient to apply a Bayesian method to simplify the problem of ML estimates. Then we need assumptions to independent prior distributions for $\boldsymbol{\pi}$ and $(\boldsymbol{\mu}, \boldsymbol{\Omega})$.

For the general location model, a Dirichlet prior, which is a conjugate prior for the multinomial distribution, can be applied to the cell probabilities,

$$p(\boldsymbol{\pi}) \propto D(\boldsymbol{\alpha})$$

where $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_C\}$ is a vector of hyperparameters that can be specified before estimation.

Noninformative priors can be used for $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$. If an uniform prior is applied to $\boldsymbol{\mu}$ and a standard noninformative prior to the covariance matrix $\boldsymbol{\Omega}$, then

$$P(\boldsymbol{\mu}, \boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{-\frac{(w+1)}{2}}.$$

Hence, the posterior distribution represents the product of independent multivariate normal distributions for $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C$ given $\boldsymbol{\Omega}$ and an inverted-Wishart distribution (W^{-1}) for $\boldsymbol{\Omega}$. Moreover, a multivariate normal distribution for $\boldsymbol{\mu}$ and an inverted-Wishart distribution for covariance matrix $\boldsymbol{\Omega}$ can be applied as informative priors as well. Further discussions about prior information are shown in Schafer [18].

3.3.1.4 Multiple imputation for missing categorical and continuous covariates

When there are missing values in both categorical variables $\mathbf{Q}$ and continuous variables $\mathbf{Y}$, the predictive distribution of the missing data given the observed data should be considered. Let $\mathbf{q}_{i,obs}$ and $\mathbf{q}_{i,mis}$ represent vectors of the values of observed and missing parts of categorical variables for subject i, and $\mathbf{y}_{i,obs}$ and $\mathbf{y}_{i,mis}$ be vectors of the values of observed and missing parts of continuous variables for subject i. $\mathbf{O}_i(q)$ denotes the subset corresponding to the observed parts of the categorical covariates, and $\mathbf{M}_i(q)$ denotes the subset corresponding to the missing parts of the categorical covariates.

The predictive probability of falling cell q given the observed data is

$$P(\mathbf{u}_i = \mathbf{E}_q | \mathbf{q}_{i,obs}, \mathbf{y}_{i,obs}, \boldsymbol{\theta}) = \frac{\exp(\xi^*_{q,i})}{\sum_{\mathbf{M}_i(q)} \exp(\xi^*_{q,i})} \tag{3.8}$$

over the cells q that $\mathbf{q}_{i,obs} \in \mathbf{O}_i(q)$. It is also the posterior probability that subject i falls into cell q given the observed continuous variables in $\mathbf{y}_{i,obs}$, and the subject is restricted to be in one of the cells in the contingency table. $\xi^*_{q,i}$ is the linear discriminant function of $\mathbf{y}_{i,obs}$ with respect to $\mu_{q,i,obs}$, which is

$$\xi^*_{q,i} = -\frac{1}{2}\boldsymbol{\mu}_{q,i,obs}{}^T\boldsymbol{\Omega}_{q,i,obs}{}^{-1}\boldsymbol{\mu}_{q,i,obs} + \sum_{j\in\mathbf{O}_i(y)} \mu_{q,j,obs}y_{ij} + \log\pi_q \tag{3.9}$$

where $\boldsymbol{\mu}_{q,i,obs}$ and $\boldsymbol{\Omega}_{q,i,obs}$ are the subvector of mean and submatrix of covariance in cell q of the continuous variables $\mathbf{y}_{i,obs}$ for subject i, $\mu_{q,j,obs}$ is the $(q,j)^{th}$ element of $\boldsymbol{\mu}_{i,obs}$, and $\mathbf{O}_i(y)$ is the subset of $\{1,\ldots,q\}$ corresponding to the variables in $\mathbf{y}_{i,obs}$.

The discriminant $\xi^*_{q,i}$ and the parameters of the multivariate regression of $\mathbf{y}_{i,mis}$ on $\mathbf{y}_{i,obs}$ can be obtained by a single application of the sweep operator [32]. Let the parameters of the general location model be arranged into a matrix,

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{\mu}^T \\ \boldsymbol{\mu} & \mathbf{P} \end{bmatrix},$$

where $\mathbf{P}$ is a $C \times C$ matrix with elements $p_q = 2\log\pi_q$ on the diagonal and zeros elsewhere. If we sweep this $\boldsymbol{\theta}$-matrix on the positions in $\boldsymbol{\Omega}$ corresponding to $\mathbf{y}_{i,obs}$, then a transformed version of the parameters can be written as

$$\boldsymbol{\theta}_{obs} = \begin{bmatrix} \boldsymbol{\Omega}_{obs} & \boldsymbol{\mu}^T_{obs} \\ \boldsymbol{\mu}_{obs} & \mathbf{P}_{obs} \end{bmatrix}.$$

Where $p_{q,obs} = -\boldsymbol{\mu}^T_{q,i,obs}\boldsymbol{\Omega}^{-1}_{q,i,obs}\boldsymbol{\mu}_{q,i,obs} + 2\log\pi_q$ are the diagonal elements of $\mathbf{P}_{obs}$ corresponding to cell q.

EM Algorithm: to compute maximum-likelihood estimates (MLE) for the parameters of the general location model from missing mixed-type data, then the MLEs obtained from EM-algorithm can be used as starting values of parameters for data augmentation.

Notice that the likelihood is a linear function of the sufficient statistics $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{S}_3$, so that the EM algorithm can be used to obtain the MLEs of the general location model. The MLEs of parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Omega})$ are given by

$$\hat{\boldsymbol{\pi}} = n^{-1}\mathbf{x} \tag{3.10}$$

$$\hat{\boldsymbol{\mu}} = \mathbf{S}_3^{-1}\mathbf{S}_2 \tag{3.11}$$

$$\hat{\boldsymbol{\Omega}} = n^{-1}(\mathbf{S}_1 - \mathbf{S}_2^T\mathbf{S}_3^{-1}\mathbf{S}_2) \tag{3.12}$$

The E-step: Find the expectations of sufficient statistics given the observed components of covariates $\mathbf{z}_{i,obs}$ and the current values of $\boldsymbol{\theta}^t = (\boldsymbol{\pi}^t, \boldsymbol{\mu}^t, \boldsymbol{\Omega}^t)$ by sweeping on $\boldsymbol{\theta}^t$. Three steps are conducted:

Step 1. Find the expectation of $\mathbf{S}_3$. Note that the expectations of $\mathbf{u}_i$ are the predictive probabilities in (3.8):

a. Sweep the parameter matrix $\boldsymbol{\theta}^t$ on positions corresponding to $\mathbf{y}_{i,obs}$ to get $\boldsymbol{\theta}^t_{obs}$.

b. Calculate the discriminant $\xi^*_{q,i}$ (3.9) for all cells q for which $\mathbf{q}_{i,obs} \in \mathbf{O}_i(q)$ given $\mathbf{y}_{i,obs}$ and $\boldsymbol{\theta}^t_{obs}$.

c. Normalized the terms $\exp(\xi^*_{q,i})$ for these cells to obtain the predictive probabilities

$$\pi_{q,i,obs} = \frac{\exp(\xi^*_{q,i})}{\sum_{M_i(q)} \exp(\xi^*_{q,i})} \tag{3.13}$$

Step 2. Obtain the expectation of $\mathbf{S}_2$ based on the predictive probabilities: set

$$E(\mathbf{u}_{q,i}\mathbf{y}_i|\mathbf{Z}_{obs}, \boldsymbol{\theta}) = \pi_{q,i,obs}\mathbf{y}_{q,i,obs} \tag{3.14}$$

if the unit i possibly belongs to cell q and 0 otherwise, where $\mathbf{y}_{q,i,obs}$ is the predict mean of $\mathbf{y}_i$ given $\mathbf{y}_{i,obs}$ and given that subject i falls into cell q. Then, for $j, k = 1, \ldots, w$,

- if $j \in \mathbf{O}_i(y)$, then $y_{q,ij,obs} = y_{ij}$;
- if $j \in \mathbf{M}_i(y)$, then $y_{q,ij,obs} = E(y_{ij}|\mathbf{y}_{i,obs}, \mathbf{u}_i = \mathbf{E}_q) = \mu_{q,j,obs} + \sum_{k\in\mathbf{O}_i(y)} \sigma_{jk,obs}y_{ik}$ are the predicted values from the multivariate regression of $\mathbf{y}_{i,mis}$ on $\mathbf{y}_{i,obs}$ within cell q, that is, $y_{q,ij,obs}$ is obtained from the regression of y_{ij} on variables in $y_{i,obs}$ in cell q. $\sigma_{jk,obs}$ is the (j,k)th element of $\boldsymbol{\Omega}_{obs}$.

Step 3. Find the expectation of $\mathbf{S}_1$:

$$E(y_{ij}y_{ik}|\mathbf{Z}_{obs},\boldsymbol{\theta}) = \sum_{\mathbf{M}_i(q)} \pi_{q,i,obs} E(y_{ij}y_{ik}|\mathbf{Z}_{obs},\boldsymbol{\theta},u_{q,i}=1), \tag{3.15}$$

where the sum is taken over all cells q that $\mathbf{q}_{i,obs} \in \mathbf{O}_i(q)$, and $j,k = 1,\ldots,w$, then

$$E(y_{ij}y_{ik}|\mathbf{Z}_{obs},\boldsymbol{\theta},u_{q,i}=1) = \begin{cases} y_{ij}y_{ik} & \text{if both } y_{ij} \text{ and } y_{ik} \text{ observed,} \\ y_{ij}y_{q,ik,obs} & \text{if } y_{ik} \text{ is missing,} \\ y_{q,ij,obs}y_{q,ik,obs} + \sigma_{jk,obs} & \text{if both are missing.} \end{cases}$$

Moreover, the subjects $i = 1,\ldots,n$ in the dataset are cycled through in the E-step, and sweeping $\boldsymbol{\theta}$ on the positions corresponding to $\mathbf{y}_{i,obs}$ and summing the contributions (3.13)-(3.15) of subject i to the expectations of the sufficient statistics.

The M-step: After obtaining the expectations of the sufficient statistics given observed component of variables and $\boldsymbol{\theta}^t$ in E-step, the M-step preforms by using (3.10)-(3.12) to compute updated estimate of $\boldsymbol{\theta}^{t+1}$.

Iterate these steps until convergence.

Data Augmentation: A Markov Chain Monte Carlo method for generating posterior draws the parameters of a general location model, given a matrix of missing mixed-type data. In the I-step, the subjects $i = 1,\ldots,n$ in the dataset are cycled through , $\boldsymbol{\theta}$ is swept to obtain the parameters of the predictive distribution of the missing variables given the observed variables; then the values are drawn from the predictive distribution to impute missing variables and compute complete-data sufficient statistics $(\mathbf{S}_1,\mathbf{S}_2,\mathbf{S}_3)$. After getting complete data, in P-step, a new value of $\boldsymbol{\theta}$ is draw from its posterior distribution given complete data from I-step.

The I-step, which includes two steps, is to draw missing data $\mathbf{z}_{i,mis}^{t+1}$ for subject i from the predictive distribution of $\mathbf{z}_{i,mis}$ given $\mathbf{z}_{i,obs}$ and the current estimates of $\boldsymbol{\theta}^t$.

Step 1. Draw $\mathbf{u}_i^{t+1}$ from predictive distribution given $\mathbf{q}_{i,obs} \in \mathbf{O}_i(q)$, which is a multinomial distribution with cell probabilities given by (3.8).

Step 2. Draw $\mathbf{y}_{i,mis}^{t+1}$ given $\mathbf{u}_i^{t+1}$ and $\mathbf{y}_{i,obs}$ based on the multivariate regression $\mathbf{y}_{i,mis}$ regression on $\mathbf{y}_{i,obs}$. The conditional distribution of $\mathbf{y}_{i,mis}$ given $\mathbf{y}_{i,obs}$, $\mathbf{u}_i$ and $\boldsymbol{\theta}^t$ is the multivariate normal with means

$$y_{q,ij,obs} = \mu_{q,j,obs} + \sum_{k \in O_i(y)} \sigma_{jk,obs} y_{ik},$$

The covariances can be simulated by Cholesky factorization discussed in Schafer [18]. Therefore, $y_{q,ij,obs}$ is the simulated draw of $\mathbf{y}_{i,mis}^{t+1}$.

The P-step, which includes three steps, is to draw estimate $\boldsymbol{\theta}^{t+1} = (\boldsymbol{\pi}^{t+1}, \boldsymbol{\Omega}^{t+1}, \boldsymbol{\mu}^{t+1})$ from their posterior distributions, given the complete versions of $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$ from the I-step. Under a noninformative prior distribution

$$P(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Omega}) \propto \left(\prod_q \pi_q^{(\alpha_q - 1)} \right) |\boldsymbol{\Omega}|^{(\frac{w+1}{2})} .$$

Therefore, the posterior distributions of parameters are

$$\begin{aligned} \boldsymbol{\pi}|\mathbf{Z} &\propto D(\boldsymbol{\alpha} + \mathbf{x}), \\ \boldsymbol{\Omega}|\boldsymbol{\pi}, \mathbf{Z} &\propto W^{-1}(n - C, (\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}})^{-1}), \\ \boldsymbol{\mu}_q|\boldsymbol{\pi}, \boldsymbol{\Omega}, \mathbf{Z} &\propto N(\hat{\boldsymbol{\mu}}_q, x_q^{-1} \boldsymbol{\Omega}), \end{aligned}$$

Step 1. Draw π_q^{t+1} for each cell q from the standard gamma distribution with shape parameters $x_q + \boldsymbol{\alpha}_q$, where $\boldsymbol{\alpha} = \{\alpha_q\}$ is an array of hyperparameters that can be specified beforehand; x_q is the q diagonal element of $\mathbf{S}_3$, and $\sum_q \pi_q = 1$.

Step 2. Draw $\boldsymbol{\Omega}^{t+1}$ from an inverted-Wishart distribution with parameters $(n - C)$ and $(\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}})^{-1}$, where $\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} = \mathbf{S}_1 - \mathbf{S}_2^T \mathbf{S}_3^{-1} \mathbf{S}_2$.

Step 3. Draw $\boldsymbol{\mu}_q^{t+1}$ from normal distribution with mean $\hat{\boldsymbol{\mu}}_q = \mathbf{S}_3^{-1} \mathbf{S}_2$ and variance $x_q^{-1} \boldsymbol{\Omega}$.

Iterate until convergence.

When sample size is relative small compared with the total number of cells, then the Bayesian Iterative Proportional Fitting (BIPF) algorithm [18] for the general location model can be used to reduce estimated parameters. In the mix library, data augmentation for the general location model is indicated by da.mix, and BIPF algorithm for the restricted general location model is indicated by dabipf.mix.

Imputations of missing covariates are created by the I–step of the data augmentation, and the imputations substitute the missingness, then the dataset is complete. Multiple imputations (imp.mix) are obtained by repeating data augmentation m times, so we have m sets of complete data. MIX package for software R, which was originally designed by Schafer [18], can be downloaded from CRAN.

3.4 INTRODUCTION TO MICE

The basic method of Multiple Imputation by Chained Equations (MICE) is called compound imputation method [37]. The details of this method are outlined below.

3.4.1 Variable-by-variable Gibbs sampling algorithm

Let $\mathbf{z} = \{z_1, \ldots, z_k\}$ be the variables in a Cox proportional hazards model, and let $\mathbf{z}_{mis} = \{z_{j,mis}\}$ denote a l dimensional vector of variables that have missing values, and $\mathbf{z}_{obs} = \{z_{j,obs}\}$ denote a p dimensional vector of variables that are fully observed. A variable-by-variable version of the Gibbs sampling algorithm is applied to draw imputations for the j^{th} missing variable, $z_{j,mis}$. "Variable-by-variable" means that the underlying statistical model is specified by separated imputation models, each of which represents a statistical relationship between an incomplete variable, $z_{j,mis}$, and a set of predictor variables, $\mathbf{z}_{obs}$. An advantage of the variable-by-variable Gibbs sampling algorithm is that only the relevant predictor variables for the incomplete variable are included in the corresponding imputation models, thus reducing the number of parameters.

Let $Z^*_{j,mis}$ be the imputation of $z_{j,mis}$, and $\boldsymbol{\theta}_j$ denote unknown parameters in an imputation model. Therefore, at an iteration t, the Gibbs sampling algorithm for generating $\mathbf{Z}^t_{mis}$ from $\mathbf{Z}^{t-1}_{mis}$ can be written as

$$\begin{aligned}
\boldsymbol{\theta}^t_1 &\sim P(\boldsymbol{\theta}_1|[Z^{t-1}_1,\ldots,Z^{t-1}_l,\mathbf{z}_{obs}]) \\
Z^t_{1,mis} &\sim P(z_{1,mis}|[Z^{t-1}_1,\ldots,Z^{t-1}_l,\mathbf{z}_{obs},\boldsymbol{\theta}^t_1]) \\
&\vdots \\
\boldsymbol{\theta}^t_j &\sim P(\boldsymbol{\theta}_j|[Z^t_1,\ldots,Z^t_{j-1},Z^{t-1}_j,\ldots,Z^{t-1}_l,\mathbf{z}_{obs}]) \\
Z^t_{j,mis} &\sim P(z_{j,mis}|[Z^t_1,\ldots,Z^t_{j-1},Z^{t-1}_j,\ldots,Z^{t-1}_l,\mathbf{z}_{obs},\boldsymbol{\theta}^t_j]) \\
&\vdots \\
\boldsymbol{\theta}^t_l &\sim P(\boldsymbol{\theta}_l|[Z^t_1,\ldots,Z^t_{l-1},Z^{t-1}_l,\mathbf{z}_{obs}]) \\
Z^t_{l,mis} &\sim P(z_{l,mis}|[Z^t_1,\ldots,Z^t_{l-1},Z^{t-1}_l,\mathbf{z}_{obs},\boldsymbol{\theta}^t_l])
\end{aligned} \tag{3.16}$$

3.4.2 Elementary imputation method

For Equation (3.16), the imputations, $Z^t_{j,mis}$, for $z_{j,mis}$ at an iteration t are generated according to different imputation models. The model that generates imputations for the j^{th} variable $z_{j,mis}$ given fully observed variables $\mathbf{z}_{obs}$ is called an *elementary imputation method*. *A compound imputation method* can be defined as the imputation methods that generate imputations for more than one incomplete variable.

The standard imputation method for a binary variable uses a logistic regression model. Let $Z^*_{j,mis}$ be a binary imputation variable with $\pi = P(Z^*_{j,mis} = 1|\mathbf{z}_{obs})$, then logistic regression model is given by

$$\ln\left(\frac{\pi}{1-\pi}\right) = \theta_0 + \theta_1 z_{1,obs} + \ldots + \theta_p z_{p,obs},$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_p)$. An imputation $Z^*_{j,mis}$ can be formulated as follows:

- Step 1. Draw $\boldsymbol{\theta}^* = (\theta^*_0, \theta^*_1, \ldots, \theta^*_p)$ from the approximate posterior distribution, which is $MVN(\hat{\boldsymbol{\theta}}, \boldsymbol{\Omega}(\hat{\boldsymbol{\theta}}))$;
- Step 2. Calculate $\pi^*_j = 1/(1 + \exp(-(\boldsymbol{\theta}^*\mathbf{z}_{obs})))$;

- Step 3. Impute a value, $Z^*_{j,mis}$ so that $Z^*_{j,mis} = 1$ with probability π^*_j and $Z^*_{j,mis} = 0$ with probability $1 - \pi^*_j$, for $j = 1, \ldots, n_{mis}$.

The standard imputation method for a categorical variable uses the polytomous regression imputation. If most of the $\mathbf{z}_{obs}$ are continuous variables, discriminant imputation is an alternative method for logistic or polytomous regression imputations. For continuous $\mathbf{z}_{mis}$, a linear regression imputation is used as a standard method, which is written as

$$z_{j,mis} = \theta_0 + \theta_1 z_{1,obs} + \ldots + \theta_p z_{p,obs} + \epsilon, \ \epsilon \sim N(0, \sigma^2),$$

and the linear regression imputation can be adjusted for a non-linear relationship between $\mathbf{z}_{mis}$ and $\mathbf{z}_{obs}$ by using transformations for one or more variables.

3.4.3 Selecting an imputation method

Two steps are proposed to find an imputation method that efficiently describes the relationship between variables and properly fits the data. Step 1 is to select relevant predictor variables for each incomplete variable $z_{j,mis}$, and step 2 is to select an elementary imputation method for each $z_{j,mis}$.

In the mice library [36] implemented in R, users can pick an elementary imputation method for each incomplete variable. Built-in elementary imputation methods are available in the mice library, such as logistic regression (impute.logreg) and polytomous logistic regression (impute.polyreg) for categorical variables; predictive mean matching (impute.pmm), Bayesian linear regression (impute.norm) and unconditional mean imputation (impute.mean) for continuous variables. In addition, users are allowed to write their own elementary imputation functions from within the Gibbs sampling algorithm, which is flexible for carrying out the complex analyses based on different missing data mechnisams. The mice library can be downloaded from CRAN directly or from the address `http://www.multipleimputation.com`.

3.5 INFERENCE BASED ON MULTIPLE IMPUTATION

The inference associated with multiple imputation was outlined by Little and Rubin [4]. Suppose that we have m complete datasets $(\mathbf{Z}_{obs}, \mathbf{Z}^i_{mis})$, $i = 1, \ldots, m$ after missing values have been filled in m times. Therefore, we can analyze each of them by fitting a specified model as if there were no missing values. In our case, we use a Cox proportional hazards model, which can be done using in standard software. Let $\hat{\boldsymbol{\theta}}_i$ and $\hat{\mathbf{V}}_i$ denote a complete-data estimate of $\boldsymbol{\theta}$ and its associated estimated variance under a certain model, respectively, for $i = 1, \ldots, m$. Then, the multiple imputation estimate of $\boldsymbol{\theta}$ is given by

$$\bar{\boldsymbol{\theta}} = \frac{1}{m} \sum_{i=1}^{m} \hat{\boldsymbol{\theta}}_i \ , \tag{3.17}$$

and the estimate of the variance is given by two components of variability: one is the average within-imputation variance, which is

$$\bar{\mathbf{V}} = \frac{1}{m} \sum_{i=1}^{m} \hat{\mathbf{V}}_i \ ; \tag{3.18}$$

and the other is the between-imputation variance, which is

$$\mathbf{B}_v = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}})^2 \ . \tag{3.19}$$

Then, the total variance associated with $\bar{\boldsymbol{\theta}}$ is given by

$$\boldsymbol{\Sigma} = \bar{\mathbf{V}} + \left(1 + \frac{1}{m}\right) \mathbf{B}_v \ , \tag{3.20}$$

where $(1 + \frac{1}{m})$ is an adjustment for a finite m. Moreover, an estimate of the fraction of information about $\boldsymbol{\theta}$ missing due to nonresponse is defined as

$$\hat{\boldsymbol{\gamma}} = (1 + 1/m)\mathbf{B}_v/\boldsymbol{\Sigma} \ . \tag{3.21}$$

In addition, for large sample size, the inference about $\boldsymbol{\theta}$ is based on

$$(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})\boldsymbol{\Sigma}^{1/2} \sim t_\nu \ , \tag{3.22}$$

with degrees of freedom

$$\nu = (m-1)\left[1 + \frac{1}{m+1}\frac{\bar{\mathbf{V}}}{\mathbf{B}_v}\right]^2 .$$

The main feature of multiple imputation is that a very small value of m will suffice, for example, $m = 3$ is often adequate, and usually 5 to 10 imputations are used in practice. The reasons are twofold. First, multiple imputation depends on simulations to solve only the missing-data problem. Thus, choosing a large m would result in an unimportant gain in efficiency, and unless rates of missingness are very high, there tends to be no real benefit in using more than 5 to 10 imputations. Secondly, the reason why valid inferences can be obtained with very small m is that the rules for combining the m complete-data analyses explicitly account for Monte Carlo errors. Both the point and variance estimates contain a predictable amount of simulation errors because of the finite m, and the width of the interval is accordingly adjusted to maintain the appropriate probability of convergence [18].

3.6 APPLICATION TO SURVIVAL ANALYSIS WITH MISSING COVARIATES

For the simulations, we generated data using a Cox proportional hazards model given by

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{z}) = 1 \times \exp(\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4 + \beta_5 z_5)$$

where the baseline failure times were generated from an exponential distribution with parameter, $\lambda_0(t) = 1$, and the true coefficients were set to be zeros, that is,

$$\boldsymbol{\beta}' = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)' = (0, 0, 0, 0, 0)' .$$

Similar to the simulation strategy used by Herring and Ibrahim [10], we generate z_1, z_2 and z_3 to be continuous covariates, and z_4 and z_5 to be categorical covariates. The covariate z_1 is from a normal distribution with mean 0 and variance 0.5, z_2 was from a normal distribution with mean 3 and variance 0.5, and z_3 is from a normal distribution with mean 1 and variance 0.5. Also, z_4 is Bernoulli with success probability 0.6, and z_5 is Bernoulli with success probability $\exp(\alpha_0 + \alpha_1 z_4)/(1 + \exp(\alpha_0 + \alpha_1 z_4))$, where $(\alpha_0, \alpha_1) = (1, 1)$. The survival times

are randomly censored with probability 0.10. The covariates, z_1 and z_4, are fully observed for each subject, and z_2, z_3, and z_5 are missing for some subjects. Missing data are generated as follows: for subject i, let (r_{i2}, r_{i3}, r_{i5}) indicate whether (z_{i2}, z_{i3}, z_{i5}) are missing, and $r_{ij} = 1$ if the j^{th} covariate is observed for subject i and 0, otherwise. Therefore, the missing data mechanisms of z_2 and z_3 are

$$Pr(r_{i2} = 0|x_i(t), \mathbf{z}_{i,obs}, \boldsymbol{\phi}_2) = \frac{\exp(\phi_{20} + \phi_{21}x_i(t) + \phi_{22}z_{i1})}{1 + \exp(\phi_{20} + \phi_{21}x_i(t) + \phi_{22}z_{i1})} ,$$

$$Pr(r_{i3} = 0|x_i(t), \mathbf{z}_{i,obs}, z_{i2}, \boldsymbol{\phi}_3) = \frac{(\phi_{30} + \phi_{31}x_i(t) + \phi_{32}z_{i1} + \phi_{33}z_{i2})}{1 + (\phi_{30} + \phi_{31}x_i(t) + \phi_{32}z_{i1} + \phi_{33}z_{i2})} ,$$

respectively, where $\boldsymbol{\phi}_2 = (-2.0, 0.0, 2.5)$ and $\boldsymbol{\phi}_3 = (-1.0, 0.0, 0.5, 0.5)$. The missing data mechanism of z_5 is

$$Pr(r_{i5} = 1|x_i^*(t), \mathbf{z}_{i,obs}, \boldsymbol{\phi}_5) = \frac{\exp(\phi_{50} + \phi_{51}x_i^*(t) + \phi_{52}z_{i4} + \phi_{53}z_{i4}x_i^*(t))}{1 + \exp(\phi_{50} + \phi_{51}x_i^*(t) + \phi_{52}z_{i4}\phi_{53}z_{i4}x_i^*(t))} ,$$

where $\boldsymbol{\phi}_5 = (0.12, -1.00, 0.20, 1.50)$, and $x_i^*(t) = (x_i(t) - \mu_{x_i(t)})/\sigma_{x_i(t)}$, $X(t)$ is the observed event times. Simulation studies were conducted to examine properties of parameter estimates of multiple imputation with 1000 replications and sample sizes $n = 1000$ and $n = 200$, respectively, in each replication.

3.6.1 Simulations under MAR

In the simulations, missing data mechanism was MAR, and the missing data pattern was not monotonic. The percentage of at least one covariate missing ranges from 40% to 50%. Table 3 summarizes simulation results for four types of estimates: "full data" that estimates from the generated data before deletion of missing covariates, "complete cases" (CC) that deletes all subject with missing values, the multiple imputation method by MICE, and the multiple imputation method by MIX.

When the sample size is 1000, the means of coefficients estimated by MICE are slightly closer to full data estimators than those by MIX. Standard errors from MICE are larger than those from MIX except for z_2 where the standard errors are slightly smaller. For the sample size of 200, means of estimated coefficients by using MICE are closer to full data analysis than those by using MIX. MIX on average has smaller standard errors than MICE in all variables.

The performance of both MICE and MIX is better than that of CC. Histograms of coefficients of full data analysis, MICE and MIX are displayed in Figures 1–2. From those displays, we can see that MICE and MIX estimators are approximately normally distributed. Compared with distributions of full data estimators, distributions of MICE and MIX estimators match well for fully observed variables z_1 and z_4, and slights differences appear in filled–in variables z_2, z_3 and z_4.

Table 3: Means of Estimated Coefficients (and standard errors) under MAR

Effect	Method	$n = 1000$	$n = 200$
z_1	Full Data	0.000455(0.06717)	0.001890(0.15675)
	CC	0.001704(0.09566)	-0.004998(0.24160)
	MICE	0.000528(0.06742)	0.001850(0.16574)
	MIX	0.000537(0.06737)	0.002625(0.15820)
z_2	Full Data	-0.000938(0.06720)	-0.010640(0.15624)
	CC	0.001266(0.09381)	-0.016880(0.22374)
	MICE	0.000002(0.07333)	-0.010360(0.17245)
	MIX	0.000051(0.07341)	-0.009489(0.17120)
z_3	Full Data	0.001616(0.06739)	-0.004890(0.15608)
	CC	0.000952(0.09297)	-0.005880(0.22084)
	MICE	0.001930(0.07810)	-0.004284(0.18306)
	MIX	0.002478(0.07798)	-0.003370(0.18150)
z_4	Full Data	0.002396(0.06986)	0.000745(0.16072)
	CC	0.001982(0.09436)	0.001345(0.22188)
	MICE	0.002405(0.07044)	0.000695(0.16288)
	MIX	0.002443(0.06998)	0.000958(0.16210)
z_5	Full Data	0.004892(0.08949)	-0.004205(0.20698)
	CC	0.002680(0.12079)	-0.003137(0.28697)
	MICE	0.004670(0.09817)	-0.002753(0.22972)
	MIX	0.005071(0.08962)	-0.004704(0.20840)

The coverage by using two imputation methods are summarized in Table 4. The coverages estimated by MICE and MIX are larger than 95% when sample size is 1000. On the other hand, when the sample size is 200, the estimated coverages from MICE and MIX lie in the interval $95\% \pm 1.4\% = 95\% \pm 1.96\sqrt{\frac{0.95(1-0.95)}{1000}} \times 100\%$.

Table 4: Comparisons of coverages between MICE and MIX under MAR

	coverage	
Method	n=1000	n=200
MICE	96.9%	96.2%
MIX	97.2%	96.1%

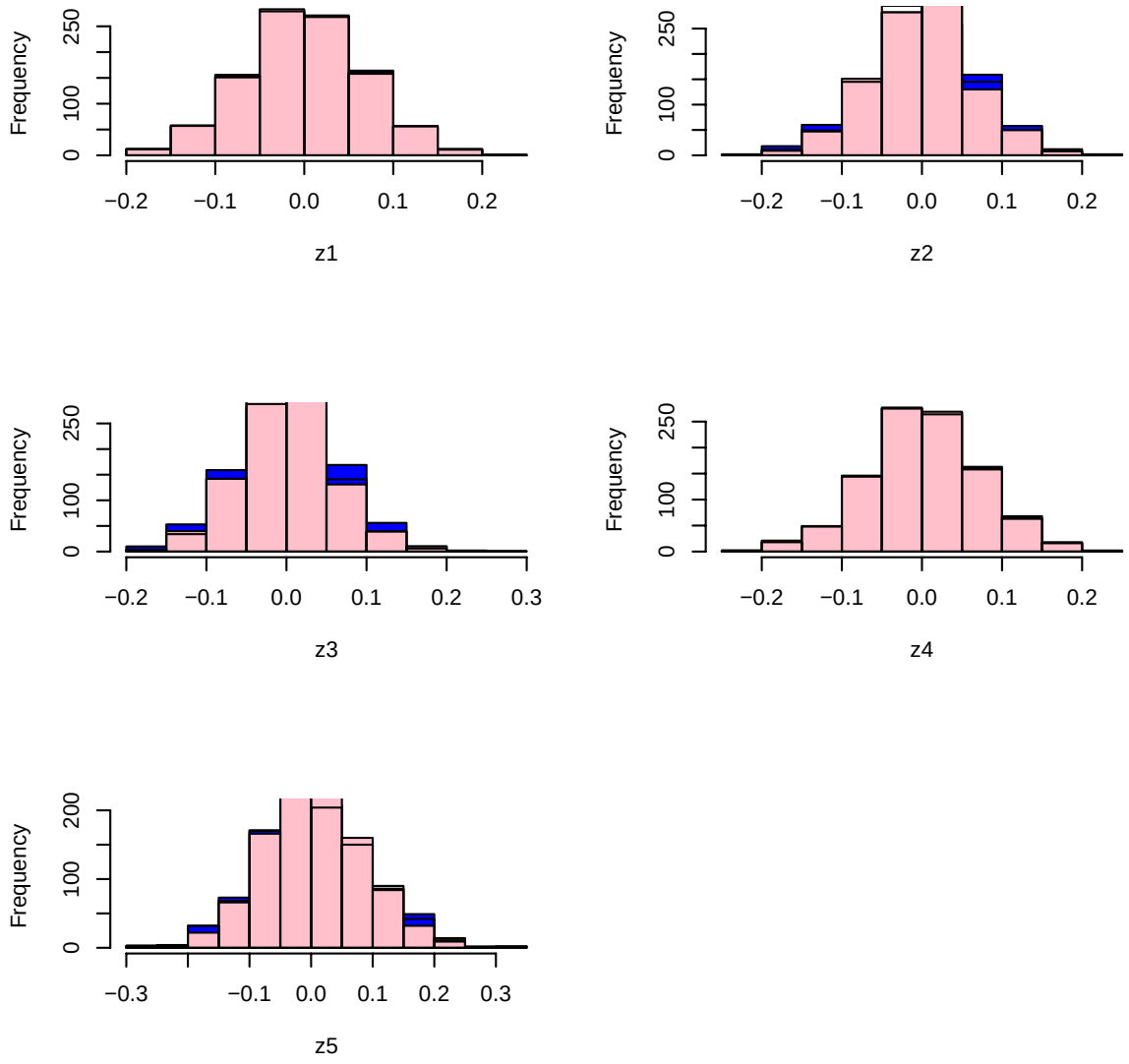
3.6.2 Simulations under NMAR

The robustness of both methods under NMAR were also considered. To do this, assume that the missingness of z_3 depends not only on the observed elements of z_2 but also on the missing elements of z_2. For both small and large sample sizes (200 and 1000), MICE and MIX have less biased estimators and smaller standard errors than CC, and MIX have smaller standard errors compared with MICE in most of cases. Means of coefficients from MICE and MIX perform similarly for the large sample size (Table 5). However, the coefficient of z_3 from MIX has a different sign compared with that from full data analysis for small sample size. When the sample size is 1000, MICE has overinflated coverages, and MIX's coverage lies in the interval $95\% \pm 1.4\%$. When sample size is 200, both MICE and MIX have reasonable estimated coverages. The estimated coverages are presented in Table 6.

3.6.3 Simulations with missing non-normal continuous covariates under MAR

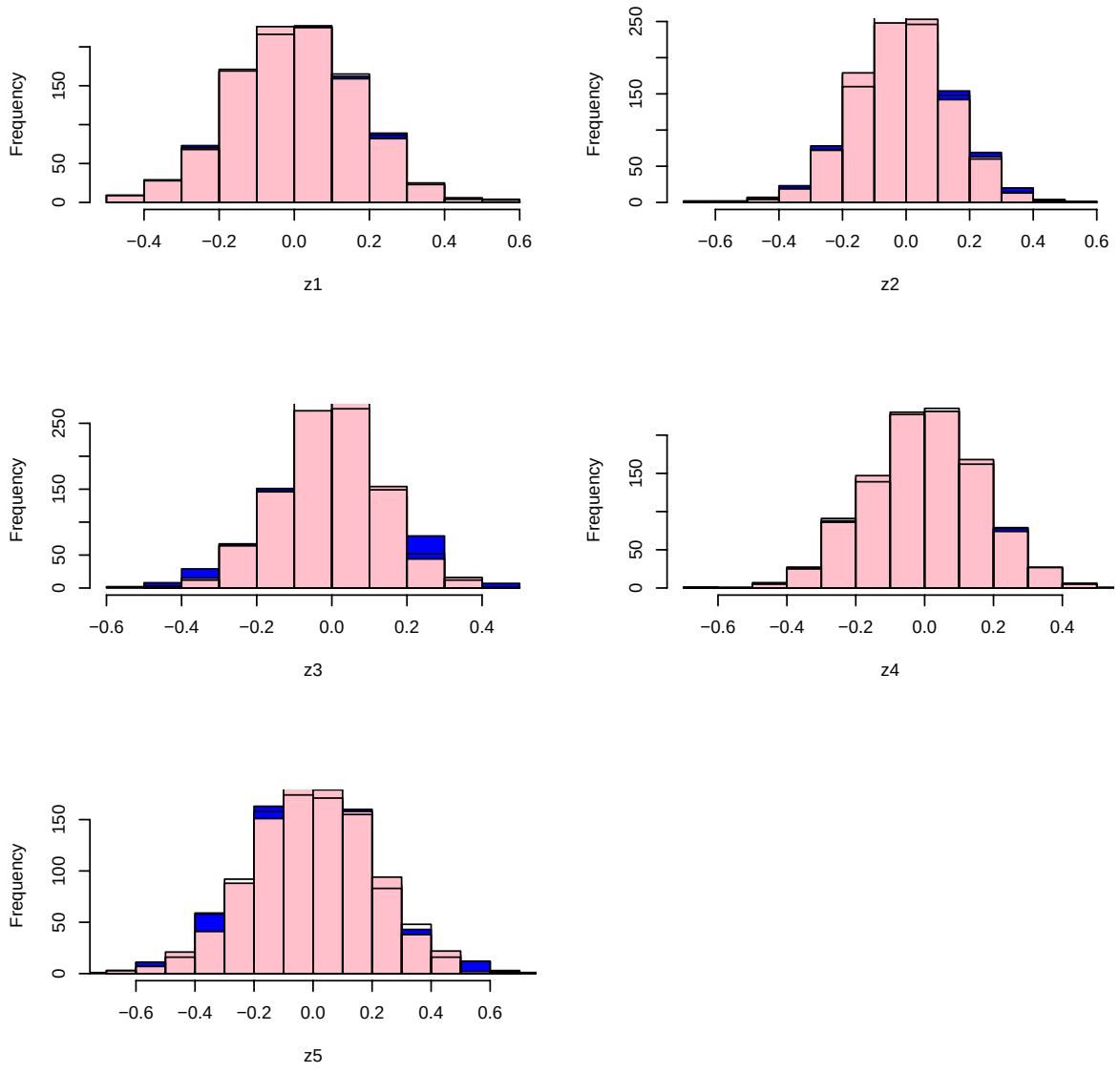
Another simulation was carried out to investigate the robustness of MICE and MIX under MAR when missing continuous covariates are not from normal distributions. We assume that all conditions are the same as the first simulation except that continuous variable z_2 is assumed from uniform distribution with parameters $(0, 1)$, and z_3 is assumed from exponential distribution with hazard 1. The simulations are also repeated 1000 times, sample sizes are 200 and 1000 respectively, and 10 imputations are conducted as well.

Estimators and standard errors displayed in Table 7 show that both MICE and MIX



Blue: Full data; Pink: MICE; White: MIX

Figure 1: Histograms of Estimated Coefficients from Full data, MICE and MIX under MAR, Sample size 1000.



Blue: Full data; Pink: MICE; White: MIX

Figure 2: Histograms of Estimated Coefficients from Full data, MICE and MIX under MAR, Sample size 200.

Table 5: Means of Estimated Coefficients (and standard errors) under NMAR

Effect	Method	$n = 1000$	$n = 200$
z_1	Full Data	-0.00195(0.06733)	0.00151(0.15576)
	CC	0.00127(0.09569)	-0.00333(0.22790)
	MICE	-0.00193(0.06751)	0.00155(0.15783)
	MIX	-0.00195(0.06747)	0.00155(0.15740)
z_2	Full Data	0.00296(0.06725)	-0.00942(0.15633)
	CC	0.00459(0.09399)	0.00072(0.22530)
	MICE	0.00285(0.07340)	-0.00422(0.17271)
	MIX	0.00283(0.07340)	-0.00399(0.17240)
z_3	Full Data	-0.00252(0.06725)	0.00244(0.15563)
	CC	-0.00222(0.09278)	0.00269(0.22161)
	MICE	-0.00142(0.07839)	0.00264(0.18526)
	MIX	-0.00119(0.07843)	-0.00023(0.18330)
z_4	Full Data	0.00040(0.06983)	0.00036(0.16084)
	CC	0.00499(0.09429)	-0.00219(0.22224)
	MICE	0.00040(0.07020)	-0.00111(0.16306)
	MIX	0.00045(0.06995)	-0.00046(0.16240)
z_5	Full Data	0.00576(0.08947)	0.00778(0.20724)
	CC	0.00851(0.12042)	0.02769(0.28782)
	MICE	0.00539(0.09822)	0.01378(0.22976)
	MIX	0.00554(0.08960)	0.00773(0.22976)

Table 6: Comparisons of coverages between MICE and MIX under NMAR

	coverage	
Method	n=1000	n=200
MICE	97.1%	96.7%
MIX	96.4%	96.6%

perform better than CC. When sample size is 1000, MIX estimators are less biased and have smaller standard errors than MICE estimators. Coefficients of categorical variables z_4 and z_5 from MICE have different signs compared with those from full data analysis. When the sample size is 200, MIX estimators are more efficient than MICE estimators except z_1, and they have less biased coefficients in z_1, z_2 and z_5. Coverages of MIX and MICE are slightly overestimated when sample size is 1000 (see Table 8). Figures 3 – 4 display that the distributions of coefficients from both MICE and MIX are approximately normal.

3.6.4 NSABP breast cancer Data

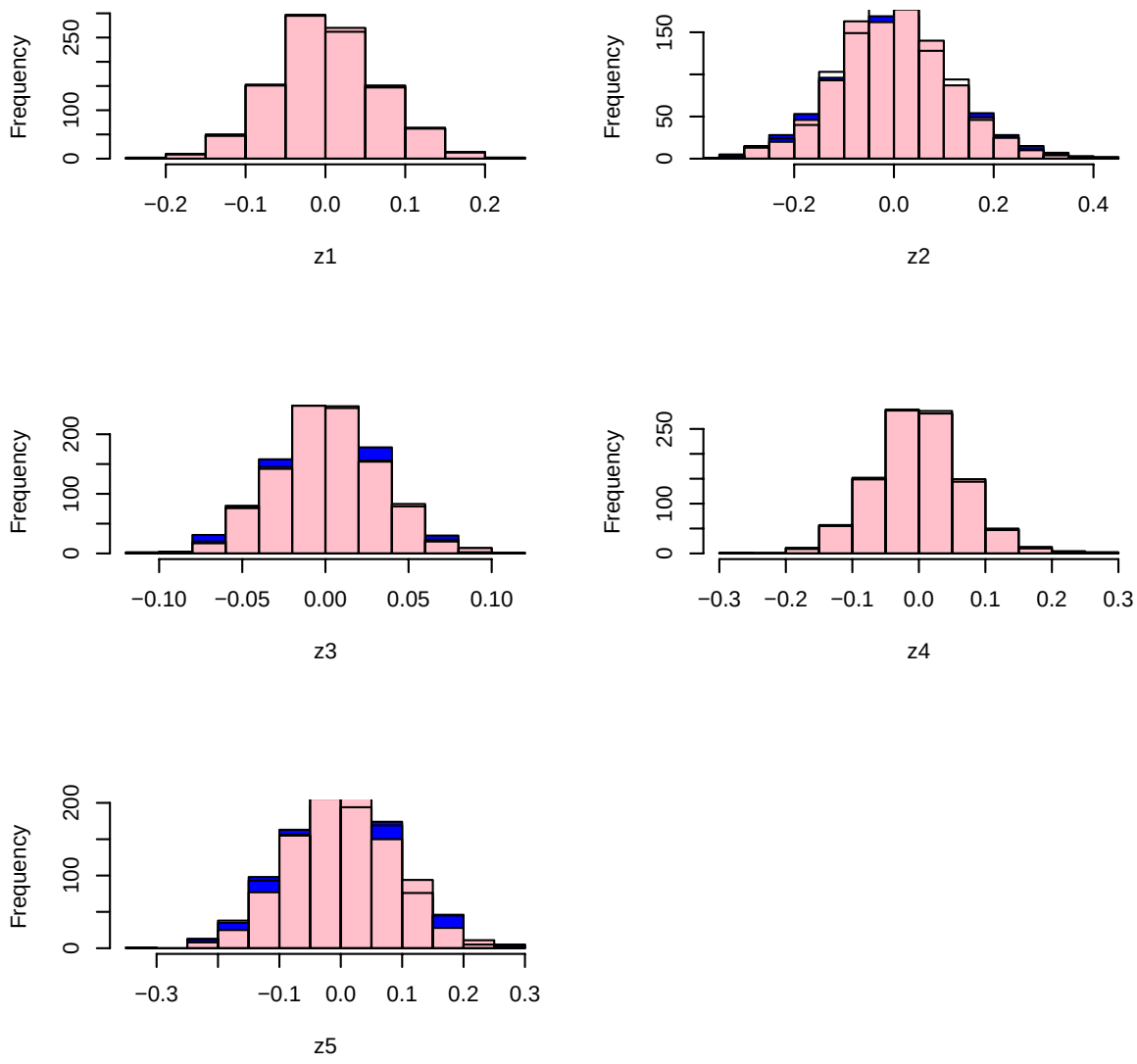
As mentioned earlier in chapter 1, the NSABP protocol B-06 is a randomized clinical trial designed to determine whether lumpectomy with or without radiation therapy was as effective as total mastectomy for the treatment of invasive breast cancer [14][15]. By the December 31, 2001, 2,163 women were entered in the study. A subset of 1,137 lumpectomy patients who had tumor free surgical margins were assessed. A Cox proportional hazards model was used to perform a multivariate analysis of variables which were significantly related to time to recurrence in the ipsilateral breast. As indicated in Table 9, two of the eight covariates listed had no missing information, another five had moderate proportions of missing information and one (Progesterone Receptor status) had a large proportion of missing information. Altogether, 49.52% of the women in this analysis had at least one missing value in their covariates. The missing data mechanism was assumed to be missing at random.

Table 7: Means of Estimated Coefficients (and standard errors) under MAR when z_2 and z_3 from non-normal distributions

Effect	Method	$n = 1000$	$n = 200$
z_1	Full Data	0.00116(0.06728)	-0.00063(0.15652)
	CC	0.00158(0.13149)	0.00881(0.31372)
	MICE	0.00109(0.06762)	-0.00109(0.15915)
	MIX	0.00112(0.06745)	-0.00010(0.16620)
z_2	Full Data	-0.00151(0.11632)	0.00873(0.26825)
	CC	0.00332(0.17958)	0.01550(0.42860)
	MICE	-0.00051(0.12724)	0.00393(0.29634)
	MIX	-0.00121(0.12710)	0.00462(0.29540)
z_3	Full Data	0.00042(0.03376)	0.00554(0.07912)
	CC	0.00075(0.04819)	0.01140(0.11790)
	MICE	0.00068(0.03734)	0.00403(0.08889)
	MIX	0.00038(0.03723)	0.00341(0.08771)
z_4	Full Data	0.00041(0.06986)	0.01242(0.16072)
	CC	-0.00033(0.09739)	0.01072(0.22980)
	MICE	-0.00014(0.07026)	0.01262(0.16270)
	MIX	0.00043(0.06998)	0.01216(0.16180)
z_5	Full Data	-0.00038(0.08946)	0.00382(0.20722)
	CC	0.00142(0.12454)	-0.00154(0.29724)
	MICE	0.00295(0.09818)	0.00221(0.22926)
	MIX	-0.00034(0.08954)	0.00419(0.20830)

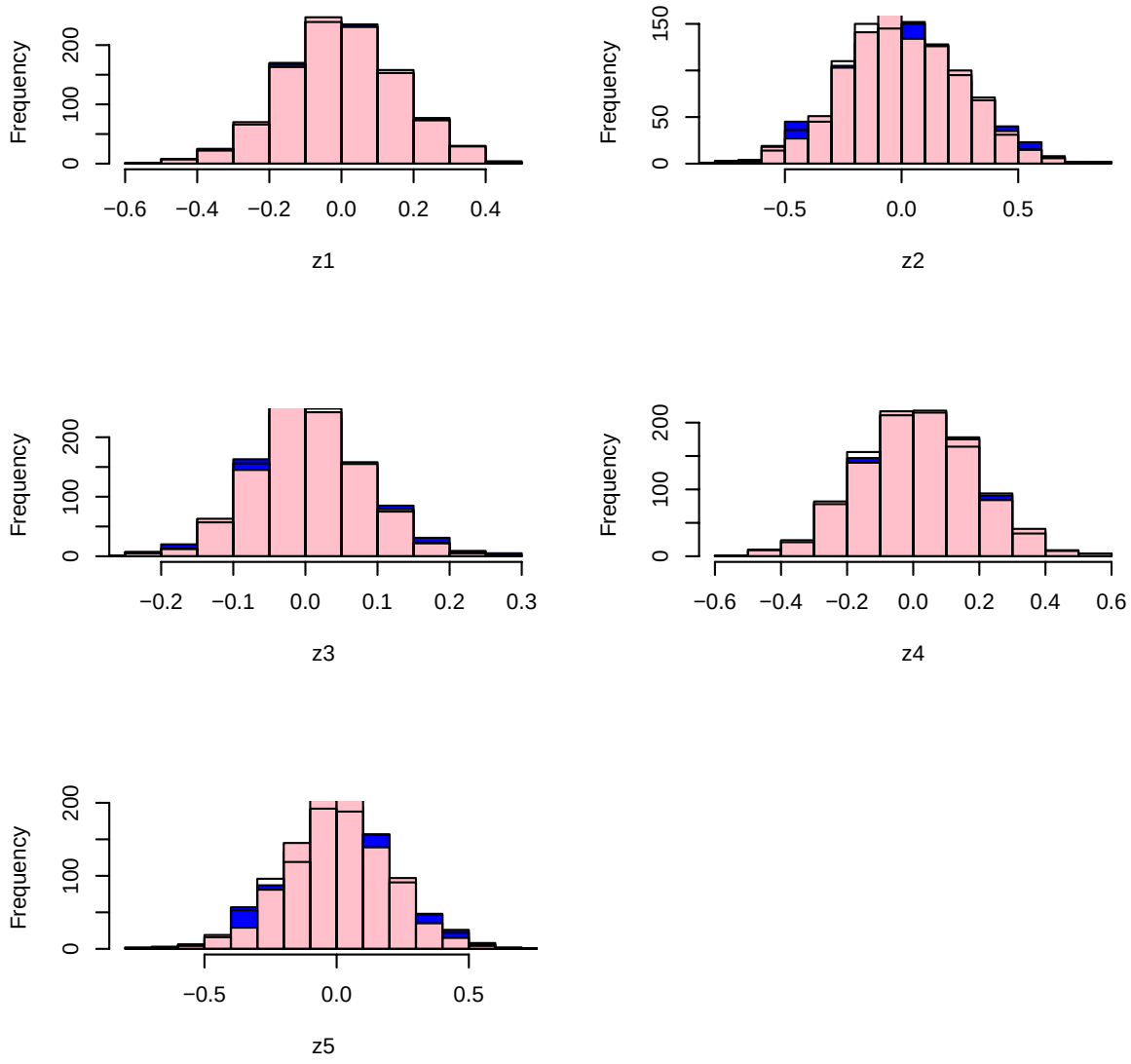
Table 8: Comparisons of coverages between MICE and MIX when z_2 and z_3 are from non-normal distributions under MAR

	coverage	
Method	n=1000	n=200
MICE	96.8%	96.4%
MIX	96.5%	95.3%



Blue: Full data; Pink: MICE; White: MIX

Figure 3: Histograms of Estimated Coefficients from Full data, MICE and MIX, Sample size 1000 for incomplete non-normal continuous covariates.



Blue: Full data; Pink: MICE; White: MIX

Figure 4: Histograms of Estimated Coefficients from Full data, MICE and MIX, Sample size 200 for incomplete non-normal continuous covariates.

Table 9: Percentage of Missing Values in Covariates

Variable	Percent Missing (%)
Age	0
Race	5.54
Tumor Type	14.16
Nodal Status	0
Nuclear Grade	11.43
Estrogen Receptor	25.24
Progesterone Receptor	36.06
Blood Vessel Invasion	14.95
At Least One Missing	49.52

Missing covariates in Cox models were treated by using three methods: the complete case analysis, and the multiple imputation methods using MICE and MIX. Variables included categorical variables race (0 = white, 1 = nonwhite), nodal statutes (0 = negative nodes, 1 = positive nodes), tumor type (0 = favorable, 1 = intermediate and 2 = unfavorable, so we created two dummy variables: Intermediate Tumor Type (Tumortype I) and Unfavorable Tumor Type (Tumortype II)), nuclear grade (0 = good, 1 = poor) and blood vessel invasion (blood) (0 = yes, 1 = no); continuous variables age (we used centered age that is age -40 and $(\text{age}-40)^2$), estrogen receptor (ER) status and progesterone receptor (PR) status. However, MIX could not run because the numbers of counts in some categories of (categorical) variables race, nodal status and blood vessel invasion were sparse. For example, the counts of the two race categories, white and nonwhite, were 1595 and 153, respectively, which resulted in empty cells when these variables were considered in combination with other categorical variables, and, hence, the program stopped running. Table 10 lists parameter estimates and standard errors in two methods, CC and MICE. Estimated coefficients using CC had different results from those using MICE. The CC method tended to have much higher values

for the coefficients of race, tumor type, nuclear grade, $(\text{age}-40)^2$ and PR, and lower values of the other covariates. Moreover, MICE estimators had smaller standard errors than CC estimators. Results from MICE indicated that race, nodal status, tumor type, nuclear grade and $(\text{age}-40)^2$ were significant at an α level of 0.05, but those from CC showed an additional significant coefficient, i.e., the one associated with (age−40). The 95% confidence intervals for the covariates are listed in Table 10.

3.7 DISCUSSION

According to a review by Burton and Altman [39], missing covariate data appeared in 81% of articles published in seven cancer journals in 2002 that involved survival analysis to evaluate potential prognostic factors. Most of the articles with missing covariate information used complete case analyses to deal with missingness. There are many reasons to do so; one of them is due to the convenience of computation. However, a complete case analysis may introduce bias depending on the missing data mechanism and the proportion of missingness. Even in cases where the covariate data are MCAR, when the bias is not an issue, the complete case analysis is inefficient. The MICE and MIX multiple imputation methods are implemented in easy-to-use software packages, which greatly simplify the computational burden while also generating less bias and more efficient estimators.

However, there are some limitations in these two methods. MIX does not always work for sparse data, that is, where many cell counts are zeros. In fact, it requires sufficient amounts of cell counts in each category to avoid empty cells in the multiple imputation, which may restrict the use of this package in practice. This is illustrated in the NSABP B-06 data. For example, since the proportion of African–Amercians was quite small, MIX failed to run due to empty cells during data augmentation. On the other hand, in theory, the restricted general location model of MIX, that is, the BIPF algorithm, can be applied to generate imputations for the sparse data or when the sample size is not large compared to the total number of cells. However, it may not be convenient to define marginal and design matrices during programming for categorical covariates with moderate numbers in each category. For

Table 10: Estimates of Covariates (and standard errors) using Complete case model and Multiple Imputation by MICE. (10 imputations were used)

Factors	CC ($n = 574$)	MICE ($n = 1137$)	95% CI_{MICE}
Race	0.472(0.200)	0.390(0.147)	(0.101, 0.679)
Nodal Status	0.462(0.123)	0.517(0.091)	(0.338, 0.696)
Tumortype I	0.484(0.224)	0.400(0.164)	(0.078, 0.723)
Tumortype II	0.721(0.233)	0.570(0.171)	(0.236, 0.905)
Nuclear Grade	0.447(0.132)	0.357(0.101)	(0.159, 0.555)
Age−40	-0.029(0.012)	-0.012(0.0091)	(-0.030, 0.006)
$(\text{Age}-40)^2$	0.0021(0.0005)	0.0015(0.00036)	(0.0008, 0.0023)
ER	-0.000029(0.000046)	-0.000024(0.00003)	(-0.000084, 0.000036)
PR	0.000015(0.000020)	0.000011(0.000015)	(-0.000019, 0.000042)
Blood	-0.870(0.419)	-0.557(0.315)	(-1.176, 0.063)

example, if a data set with n subjects has 4 categorical covariates with 4, 3, 2 and 3 categories respectively, then the design matrix will have a dimension of $n \times (1+4\times 3\times 2\times 3) = n \times 73$. Also, we could not successfully implement the BIPF algorithm (daipf.mix) for the NSABP data. In fact, MIX does not appear to run when a sample size is very small, even though its BIPF algorithm is used. The smallest sample sizes that we were able to conveniently program the package was 200 for da.mix, and 150 for daipf.mix.

MICE can be easily used in data with small sample sizes (less than 100). It generates slightly less biased estimators with slightly higher standard errors than MIX when continuous covariates are normally distributed. In addition, it allows advanced users to construct their own imputation functions by additional programming. However, the random number generation associated with MICE uses a fixed seed by default, which seems to create the same imputed values over different replications. This finding is also mentioned in Horton and Lipsitz [35]. Although typically, one chooses random seeds during programming, a non-cautious user may overlook this potential problem with the default seed. Neither MICE and MIX have built–in programs that support survival regression modeling, but MIX has a function (im.inference) for pooling inference of multiple imputation for all models, which is easier than MICE (whose function of pooling inference pool can be used only within its built–in linear regression models and generalized linear models).

The results in our simulations suggest that in some cases, even when the missing data mechanism is NMAR, both packages can perform reasonably well with respect to the parameter estimators, standard errors and coverage. However, this obviously is not true in all cases. MIX is slightly more efficient and has more reasonable coverage than MICE. Still, we find that overinflated coverages appear when using the packages under either MAR or NMAR for large sample sizes. When missing continuous covariates are from non-normal distributions, MIX appears to perform better than MICE.

We note that both packages have flexibility for implementing multiple imputation procedures, and generate reasonable results for survival data. In closing, we would agree with the following cautionary note by Horton and Lipsitz [35]: "the existence of software that facilitates its use requires the analyst to be careful about the verification of assumptions, the robustness of imputation models, and the appropriateness of inferences".

4.0 LITERATURE REVIEW ON STEPWISE REGRESSION PROCEDURES

It is often necessary to estimate a regression relation between an outcome variable and a set of explanatory variables. However, the number of potential variables to be used in a regression model is often too large and a more parsimonious model may be preferred. Selection strategies, especially stepwise methods, are widely used. Stepwise methods have been available for a long time, and basically there are three techniques: forward selection (FS), stepwise selection (SS) and backward elimination (BE). Many methods for variable selection and related issues have been developed for normally distributed outcomes and investigated in an enormous literature in this area.

Clark, *et al.* [40] suggested that stepwise methods can be used as selection techniques to choose important covariates in Cox proportional hazards models. One limitation of stepwise methods is that either the backward elimination or forward selection procedures only evaluate a small number of the set of possible models. On the other hand, backward elimination, which starts with the full model, may be possibly the best of the stepwise selection strategies. Harrell, *et al.* [41] assert that uncritical applications of modeling selection techniques may result in poorly fitted models, or more likely, inaccurately predict outcomes on new data. Steyerberg, *et.al* [42] studied the influence of stepwise selection methods on estimated logistic regression coefficients in small samples. Standard errors were expressed as the precision of the estimated regression coefficients, and the estimated standard errors for each of the covariates were computed as the average of the estimated asymptotic standard errors in each logistic regression model. Moreover, functions from the **Design** library in S-Plus for logistic regression and stepwise selection were utilized during computations. Results showed that regression coefficients in a stepwise selection model in small datasets may have a considerable

bias. Wählby, *et al.* [43] evaluated the performances of stepwise model-building strategies in population pharmacokinetics and pharmacodynamics data, and observed that the selection bias can affect the results of a covariate analysis although it was small relative to the overall variability in the estimates. Ambler, *et al.* [44] investigated the predictive accuracy when various performance measures were applied during model selections under an assumption that a true model was known. They proposed a stepdown strategy to select the "best" model and compared with BE. A stepdown procedure (which is different from FS, SS, and BE) was used to approximate a linear combination of the predictors in the model. A regression on these predictors produces a perfect fit with an R^2 (squared multiple correlation) of 1. The R^2 gets lower when variables are omitted from full models. The procedure continues until R^2 is lower than a prespecified level, such as 0.95. They fit models using different R^2 values in simulations of two data sets, and compared results with different significant levels of BE. Results of their simulations suggested that the final selected models from stepdown and BE performed nearly as well as, or may be even better, than full models.

Because of increasing computational power, a bootstrap resampling method was proposed to improve selection strategy. Altman and Andersen [45] asserted that, "Bootstrapping is an appealing method for evaluating a regression model, as it allows investigation of the consistency of the inclusion of each variable in the regression model". The bootstrap analysis is often performed to help selection of a final model by (arbitrarily) entering those variables selected by more than 50 percent of the analyses of bootstrap replications. Austin and Tu [46] applied stepwise selection methods for logistic regression models to identify the predictors of an outcome, and studied the reproducibility of logistic regression models developed using stepwise selection methods. Backward elimination, forward selection and stepwise selection were applied to select important variables from candidate variables, and these processes were repeated using 1000 bootstrap samples. The variables which had been selected in each model were recorded and the results across three variable selection methods were compared. Their conclusion showed that the stepwise selection methods resulted in unstable and not reproducible models, and the selected variables in each model were sensitive to random fluctuations in the data.

Augustin, *et al.* [47] discussed two model selection strategies that account for model

selection uncertainty in two survival data sets. One strategy was to apply an approximate Bayesian Model Averaging (BMA) adapted from Bayesian model averaging method to Cox proportional hazards model. The method first performed an initial screening of variables based on each variable's inclusion frequency in bootstrap samples to omit variables that have little influence in the corresponding models, then averaged over a set of possible models using weights estimated from bootstrap resampling. The other strategy employed either an AIC or BIC to compute weights. The two strategies produced similar results and were useful when the number of potential variables in data sets was high. In addition, the paper suggested that bootstrap resampling method had the additional positive effect of reducing the number of explanatory variables and dealing with correlated variables.

In addition, many authors studied the *stability* of a chosen regression model by using bootstrap method in the Cox proportional hazards model framework. Chen and George [48] investigated the validation of Cox's proportional hazards model to characterize pediatric acute lymphocytic leukemia data by using two stages of the bootstrap method, which imitated the original population. One stage was to select important variables via a stepwise regression procedure with 100 bootstrap samples; the other is to estimate the selected variables with 400 bootstrap samples. The bootstrap results indicated that a reasonable model construction was employed and that parameter estimates compared well with the original data set. In addition, the results of this paper showed that the bootstrap resampling technique provided an easy way to assess the whole process from selection of the best model through to the validation of the model. However, the authors examined only binary variables, so further investigation of the method may be required. Altman and Andersen [45] investigated the stability of stepwise selection methods in Cox proportional hazards model from primary biliary cirrhosis clinical trial data. Model selection results of stepwise regression analysis of 100 bootstrap samples were compared with those of a stepwise selection based on the original data, and they agreed well. Sauerbrei and Schumacher [19] developed a bootstrap model selection strategy that combined the bootstrap method with stepwise methods, which was an extension of Chen and George[48]. Stepwise selection methods were applied in each bootstrap replication to identify significant variables. The number of times that each variable included in each selected model at bootstrap replications was defined

as bootstrap inclusion fractions, and prognostically important variables should have high bootstrap inclusion fractions in B bootstrap replications. They verified the bootstrap inclusion fractions as the power of the test for $\beta_j = 0$, and also accounted for the correlation structure of the variables. Moreover, a choice of a cutpoint for the percentage inclusion depended on if factors with strong relationship with the outcome (strong factors) were the only interest in the study or if weak factors should be included as well. When weak factors were considered for the next step of analysis, a low value of the cutpoint percentage would be proper. Therefore, the authors proposed two strategies to find a compromise between the percentage inclusion cutpoint and the selection level. In strategy A, the idea was that weak factors should be selected to ensure accurate prediction. In the first step, all variables which may be important can be included in the model if the bootstrap selection frequencies exceed a low level, such as 30%. Some of remaining variables should be eliminated by variables pairwise investigations of interrelationships of inclusion frequencies. The underlying principle of strategy B was for the case that only strong factors should be selected. That is, a really strong factor should be selected into the model in nearly all cases, except when there is another highly correlated covariate. In the latter situation, at least one of the two correlated factors will be selected in nearly every bootstrap replication. Hence, in the first step, all variables with a high selection frequency enter. Variables not included may enter the model in a further step if bidimensional inclusion frequencies show that one factor from a 'correlated' pair should enter the model. Furthermore, the authors suggested that "the bootstrap inclusion frequencies may lead to a more careful interpretation of the importance of a variable than does the usual standardized parameter estimate". However, in this paper, the choices of cutpoints for the percentage inclusion in a model and the choices of selection levels were somewhat arbitrary, and a larger number (at least several hundred) of bootstrap replications may be needed. Sauerbrei [49] investigated the problems of replication stability, model complexity, selection bias and an over optimistic estimate of the predictive value of a model based on the previously proposed strategies of Sauerbrei and Schumacher [19]. Backward elimination with different selection levels, cross-validation and bootstrap resampling method were used to choose a final model. Moreover, a cross-validation approach to estimate (global) shrinkage factors was extended to parameterwise shrinkage factors (PWSFs). Vari-

ables whose regression parameter estimates were not biased due to model selection should have a PWSF of about 1. Using the bootstrap method, instability in the selected model can be recognized, particularly when a complex model including several weak factors was chosen. Furthermore, variables without influence on the outcome were selected with a probability depending on the selection level. On the other hand, PWSFs suffered from the problem of too many parameters to estimate. In addition, only parameter estimates for weak factors may be biased in a predictive model.

Much previous work has shown that the results of stepwise selection procedures are usually obtained via a single model without any information about the model's stability, and, thus, may generate tremendous difference in the selection of variables. In contrast, resampling methods—the bootstrap method— allow us to investigate the stability of procedures and select variables. The bootstrap model selection strategy proposed by Sauerbrei and Schumcher [19] combines bootstrap methods with stepwise procedures while considering the degrees of relationship (strong and weak) between outcome and covariates at the same time. Their bootstrap methods are only methods that allow one to see whether there are some really 'important' variables which should be included in a model and allow one to evaluate the importance of each of the other variables conditioned on the whole set of covariates. However, their strategies had limitations, particularly the pairwise independence tests for deciding which variables should remain in the model may discard the 'important' variables. For example, in the NSABP data, clinical tumor size and nodal status are very strong factors associated with survival. Nevertheless, the chi–square test for independence showed that the two were dependent, so one of them had to be deleted from the model according to the strategies. Hence, the results from the strategies are not reasonable and unsatisfied in some situations.

Furthermore, all the methods or strategies proposed in previous work were based on the complete–case analysis. Subjects with missing values were omitted before the model selection procedures, which may be dangerous, because the complete–case analysis can give a biased view of the relationship between the predictors and the outcome.

5.0 INTRODUCTION TO A STEPWISE MODEL SELECTION STRATEGY

As Sauerber [49] and others have pointed out, adding a variable to a model may increase goodness of fit, but the complexity of the model increases, and its predictive ability could worsen. Thus, a model selection strategy is the one that not only concerns itself with which variables should be in or out of the model, but is also concerned with the improvement of the prediction accuracy after entering or deleting the variables. Moreover, when there are missing values in the covariates, the strategy should account for missing information.

Concentrating on variable selections in the Cox model, we propose a model selection strategy for data with missing covariates. The strategy has advantages of conducting model selection and evaluating the predictive accuracy of the selected submodels at the same time [53], while taking into account missing covariates. The strategy can be formulated using three steps:

- Step 1: the imputation step:

  use MICE to compute missing values, and get complete data sets;

- Step 2: the screening step for the model selection:

  a. run the bootstrap stepwise procedure to select covariates, and calculate percentage of inclusion(PI) for each covariate in bootstrap iterations;

  b. covariates with greater than 30% PI consists of a variable pool for next step;

- Step 3: the weak factors selection:

  a. fit Cox models with strong factors ($PI > 70\%$) and combination of weak factors ($30\% < PI \leq 70\%$), then calculate the Weighted Brier Score for each model.

  b. the model with the smallest Weighted Brier Score is the best model.

Thus, we briefly introduce step 1 and step 2 in sections 5.1 and 5.2 of this chapter; step 3 is introduced in more detail in chapter 6, which includes the definitions of Weighted Brier Score and weighted survival function. Simulation studies and application are conducted to assess the strategy in chapter 7.

5.1 THE IMPUTATION STEP

Previously we compared two packages in R that implement multiple imputation methods: MICE and MIX. The results show that MICE and MIX perform reasonably well with respect to the estimation of parameters, their standard errors and coverage, even under some NMAR. However, MIX is sensitive to sparse data. Hence, we will use MICE based imputation method to fill–in the missing values of the covariates in the imputation step.

5.2 THE SCREENING STEP FOR THE MODEL SELECTION

The bootstrap method was originally proposed by Efron [50]. A Bootstrap resampling technique is used to improve the stepwise selection procedures. In the survival setting, the bootstrap technique is as follows: Assume that original data are denoted by $(T, \delta, z_1, \ldots, z_p)$; and randomly take sample of size n with replacement from the original data to obtain a bootstrap sample, repeat the procedure a large number, say B, times, to obtain our bootstrap sample. The variability of estimated characteristics of the distribution can be assessed by studying the variability of the estimate across the B bootstrap samples.

The basic idea of the bootstrap stepwise model selection method is that a stepwise method (i.e. Forward Selection (FS), Stepwise Selection (SS) or Backward Elimination (BE)) is conducted to select the prognostically important variables at each bootstrap replication. In this dissertation, we use SS with a selection level corresponding to the AIC criterion. The AIC allows us to compare nonnested models and obtained parsimoninous models. Also, we define percentage of inclusion (PI) as the number of times each covariate included in the

model over total number of bootstrap replications, and record it for the model selection. Variables with no or little prognostic influence will have low PI, since we assume that each bootstrap replication is a random sample from the original patient distribution, and thus should reflect the underlying structure of the data. Consequently, the PI in the model can be viewed as a selection criterion for the prognostic importance of a variable. We select only those variables with PI's above a minimum value, say 30%. A variable with PI greater than 70% is called a *strong factor*, and a variable with PI between 70% and 30% is called a *weak factor*. If a variable with PI less than or equal to 30% indicates that the variable has no relationship with the outcome. A discussion about PI cutoff points can be found in Sauerbrei and Schumcher [19]. A Scheme of how the bootstrap stepwise procedure works is given in Table 11. The variables with PI greater than 30% will consist of a covariate pool for the next step of the model selection, the weak factors selection.

Table 11: Results of The Stepwise Selection at 10 Bootstrap Replications

Variables	1	2	3	4	5	6	7	8	9	10	PI%
z_1	√	√						√			30
z_2	√	√	√	√	√	√	√	√			80
z_3	√	√		√	√		√			√	60
z_4		√		√	√	√		√			50
z_5											0
z_6	√	√	√	√	√		√	√			70
z_7	√	√	√	√	√	√	√	√	√	√	100
z_8	√	√	√	√	√	√	√	√	√	√	100
z_9	√		√	√	√	√	√	√	√	√	90
z_{10}											0

√ indicates that the variable is included in the model selected by SS

6.0 WEAK FACTORS SELECTION AND WEIGHTED BRIER SCORE

6.1 INTRODUCTION

In model selection, a set of 'best' predictors is usually chosen from a collection of potential predictors. Standard methods, such as C_p and adjusted R^2, are used in selecting submodels and estimating predictive accuracy, but tend to be highly biased and usually result in poor selection [52]. Some literature [53]–[56] investigated model selection methods using prediction errors or similar quantities (e.g. discrepancy) as criteria. Breiman and Spector [53] proposed a resampling method for model selection using the prediction error as a criterion in regression models. For an outcome variable y_i, a covariate vector $\mathbf{u}_i$ and a parameter vector $\boldsymbol{\beta}$, $i = 1, \ldots, n$, a regression model is given by $y_i = \boldsymbol{\beta} * \mathbf{u}_i + \varepsilon_i$, with $E\varepsilon_i = 0$, $E\varepsilon_i\varepsilon_j = \sigma^2\delta_{ij}$. Then for new data $(y_i^*, \mathbf{u}_i)$, where the y_i^* is from the distribution of y_i, the prediction error is defined as

$$PE = E\|y^* - \widehat{\mu(\mathbf{u})}\|^2 \ ,$$

where $\widehat{\mu(\mathbf{u})}$ is a predictor for y regressed on $\mathbf{u}$. Backwards deletion is used to give the sequence of submodels at each bootstrap repetition. Note that, backward deletion was not used as a selection criterion here, but rather for generating submodel candidates. Also, the PE is computed for each submodel. The submodel with smallest PE is the best model. The proposed method has fairly low bias and selects dimensionalities close to the optimal selection based on true PE. However, the PE method has only been proposed in a linear regression situation and assumes that the true PE is known.

When it comes to assess predictive accuracy in a survival analysis, there is some confusion about what quantities should be predicted. Schumacher, et al [57] pointed out "ad–hoc

approaches as P–values of the logrank test, the likelihood of a corresponding Cox regression model or ROC methodology borrowed from the evaluation of diagnostic tests are commonly used that do not fully capture the specific problems arising from survival or time–to–event data and are thus of only limited value". Graf, et al. [52] presented some approaches to assess the accuracy of prediction in the survival framework. They investigated the mean square error of prediction as a measure of accuracy, but the point predictions of event–free times would lead to inaccurate and unsatisfactory results. A second proposed measure is the expected misclassification rate that can be interpreted as 'survival at time t^*' or 'failure at time t^*', or alternatively, 'diseased' or 'not diseased'. However, in a prognostic framework, it may not be adequate to label patients this way, because the future survival status at t^* cannot be determined at, say, time $t = 0$. The expected Brier score (BS), which was originally developed for predicting accuracy of weather forecasts, is a measure of accuracy based on the estimates of event–free probabilities. The expected BS has some advantages over other predictive accuracy approaches. First, it is easily to interpreted as a mean square error of the prediction of event status. Second, the estimated survival probabilities are used to predict the event status at time t^*, which means that the BS may be preferred over the misclassification rate. The reason is because the BS is used to measure average discrepancies between true event status and estimated predictive values. A low BS indicates that the prediction is very accurate for a given time period. Graf, et al. proposed a reweighting scheme to cope with censoring. Also the quantities they used did not depend on the censoring distribution asymptotically. Schumacher, et al [57] applied the BS method in a case study. The application verified that the BS had shown a valuable way to evaluating the predictive performance of prognostic classification schemes for survival data with censored observations.

The Brier score has a meaningful interpretation even if the model is wrong, which is important since all prognostic models are bound to be misspecified to some extent [57]. Also, even when misspecifying a model, BS yields efficient estimators [58]. Moreover, BS can assess a patient specific prognosis that depends on the patients covariates, which allows comparison of different regression models by looking at their prediction error. In addition, BS has a nice feature of allowing assessment of the influence of a covariate on survival.

Despite the advantages of a BS, previous literature has assumed that all variables were

fully observed, and so, the definition of a BS with missing data was not specified. Moreover, the BS has not been used as a part of the selection strategy to determine if, say, weak factors should be in or out of the model selected by the stepwise procedure. Consequently, we define a Weighted Brier Score (WBS) by using a weighted survival function so that the loss of information due to missing values is accommodated. Also the WBS can be a criterion for selecting weak factors.

6.2 DEFINITION OF THE WEIGHTED BRIER SCORE

6.2.1 The weighted survival function

As before, for subject i, $i = 1, \ldots, n$; let T_i be the time on study for the ith subject, δ_i be the event indicator with 1 if the event has occurred and 0 if the occurences of the event is right-censored and $\mathbf{z}_i(t) = (z_{i1}(t), \ldots, z_{ip}(t))$ is a vector of covariates for the ith subject at time t which may affect the survival distribution of the time to event X, where $X(t) = \min(T, C)$ is the observed event time, and C is the censoring time. Also, let $t_1 < t_2 < \ldots < t_D$ denote the distinct event times, and $R(t_j)$ is a set of all subjects at risk just prior to t_j, and define

$$W(t_j; \boldsymbol{\beta}) = \sum_{i \in R(t_j)} \exp(\sum_{h=1}^{p} \boldsymbol{\beta}_h \mathbf{z}_{ih}) \ .$$

The estimators of the baseline survival function are the product integral of the Breslow estimator of the cumulative hazard rate [59], which in the case of no tie events, is given by

$$S_0(t|0) = \prod_{t_j \leq t} \left[1 - \frac{\delta_j}{W(t_j|\boldsymbol{\beta})} \right] \ .$$

Therefore, the estimator of the survival function for a subject with a covariate vector $\mathbf{z}$ is

$$S(t|\mathbf{z}) = S_0(t)^{\exp(\boldsymbol{\beta}'\mathbf{z})} \ .$$

When there are missing values in the covariates, we can partition the vector $\mathbf{z}$ as $\mathbf{z} = (\mathbf{z}_{obs}, \mathbf{z}_{mis})$. To account for the missing values in the original data set, we propose a Weighted Survival Function (WSF) for subject i, which can be written as

$$\begin{aligned} \widehat{S}(t|\mathbf{z}) &= \widehat{S}_0(t)^{\left(\frac{r_i}{\pi_i}\right)\exp(\boldsymbol{\beta}'\mathbf{z}_{i,obs})+\left(1-\frac{r_i}{\pi_i}\right)E\{\exp[(\boldsymbol{\beta}'\mathbf{z}_{i,mis})]\}} \\ &= \widehat{S}_0(t)^{\left(\frac{r_i}{\pi_i}\right)\exp(\boldsymbol{\beta}'\mathbf{z}_{i,obs})+\left(1-\frac{r_i}{\pi_i}\right)\exp[(\boldsymbol{\beta}'\mathbf{z}_{i,imp})]} . \end{aligned} \tag{6.1}$$

The weighted baseline survival function can be written as

$$\widehat{S}_0(t) = \prod_{t_j \le t}\left[1 - \frac{\delta_j}{\widehat{W}(t_j|\boldsymbol{\beta})}\right] , \tag{6.2}$$

$$\begin{aligned} \text{where } \widehat{W}(t_j|\boldsymbol{\beta}) &= \sum_{i\in R(t_j)}\left[\left(\frac{r_i}{\pi_i}\right)\exp(\boldsymbol{\beta}'\mathbf{z}_{i,obs}) + \left(1-\frac{r_i}{\pi_i}\right)E\{\exp(\boldsymbol{\beta}'\mathbf{z}_{i,mis})\}\right] \\ &= \sum_{i\in R(t_j)}\left[\left(\frac{r_i}{\pi_i}\right)\exp(\boldsymbol{\beta}'\mathbf{z}_{i,obs}) + \left(1-\frac{r_i}{\pi_i}\right)\exp(\boldsymbol{\beta}'\mathbf{z}_{i,mip})\right] \end{aligned} \tag{6.3}$$

where $r_i = 1$ if $\mathbf{z}_i$ is fully observed, and 0 otherwise; π_i is the probability of $\mathbf{z}_i$ being fully observed for subject i, and $\hat{\pi}_i = Pr(r_i = 1|T_i, \delta_i, \mathbf{z}_{i,obs}) = \frac{\exp(-\mathbf{v}'\mathbf{m}_i)}{1+\exp(-\mathbf{v}'\mathbf{m}_i)}$, $\mathbf{v}$ is a vector of unknown parameters and $\mathbf{m}_i$ is some function of $(t_i, \delta_i, z'_{i,obs})'$ [60]; and $\mathbf{z}_{i,imp}$ is the imputation values from MICE.

Qi, et al. [62] defined weights (the inverse of selection probability π) similar with the one that we proposed here. However, the differences are twofold. First, the goals are different. Simple weighted estimators (SWE) and fully augmented weighted estimators (FAWE) in Qi, et al. were proposed to estimate parameters of the partial likelihood function in the Cox model, but our weights are used to define the WBS, which allow us to evaluate prediction accuracy of a model. Secondly, Qi, et al. used weights in the score function, but we modified these weights in the survival function to accommodate missing data.

6.2.2 The Weighted Brier Score

For subject i, $i = 1, \ldots, n$, let $G(t) = P(C > t)$ be the censoring distribution. $\widehat{G}(T_i)$ is the Kaplan-Meier estimate of $G(t)$, which can be denoted as the marginal probability of being event-free up to time T_i. At a fixed time point t^*, the weighted Brier score for no censored data [52][57], can be defined as

$$BS(t^*) = \frac{1}{n}\sum_{i=1}^{n}(I(T_i > t^*) - \widehat{S}(t^*|\mathbf{z}_i))^2 \ .$$

When there is censoring in the data, the WBS can be written as

$$BS(t^*) = \frac{1}{n}\sum_{i=1}^{n}\left\{(0 - \widehat{S}(t^*|\mathbf{z}_i))^2 I(T_i \leq t^*, \delta_i = 1)\frac{1}{\widehat{G}(T_i)} + (1 - \widehat{S}(t^*|\mathbf{z}_i))^2 I(T_i > t^*)\frac{1}{\widehat{G}(t^*)}\right\},$$

where $\widehat{S}(t^*|\mathbf{z}_i)$ is the survival probabilities from equation (6.1), and $1/\widehat{G}(t)$ is the weight to compensate the loss caused by censoring. The WBS can be decomposited as three parts according to T_i and δ_i:

I: an event occured at or before time t^* ($T_i \leq t^*$ and $\delta_i = 1$). The contribution to the WBS is $(0 - \widehat{S}(t^*|\mathbf{z}_i))^2$, and the weight $1/\widehat{G}(T_i)$ is based on the event time T_i.

II: no event was observed before time t^* ($T_i > t^*$, and $\delta_i = 1$ or $\delta_i = 0$). The contribution to the WBS is $(1 - \widehat{S}(t^*|\mathbf{z}_i))^2$, and the weight $1/\widehat{G}(t^*)$ is based on t^*.

III: a "censor" occured before t^* ($T_i \leq t^*$ and $\delta_i = 0$). The contribution to the WBS is unknown, and the weight is zero.

After the WBS is calculated for each potential model, the model with smallest WBS is the best fit (final) model.

7.0 SIMULATION AND APPLICATIONS

7.1 MIXTURES OF CONTINUOUS AND CATEGORICAL COVARIATES

Similar to Herring and Ibrahim [10], we generated both continuous and discrete variables related to outcome for our simulation studies. z_1 is a continuous variable from normal distribution $N(0,1)$, z_2 is a continuous variable from normal distribution $N(3,25)$, and z_3 is a binary variable from binomial distribution with probability of 0.6. z_1 is setup as a strong factor by using $z_1 = t + z_1 + e$, where $e \sim N(0,1)$. z_2 and z_3 are setup as weak factors by the strategies $z_2 = z_2 + t + e_2 - 3$ and $z_3 = z_3 + t^3 + 2 * e_3$, where $e_2 \sim N(2,4)$ and $e_3 \sim N(2,100)$ respectively. z_1 has roughly 21% missing values, z_2 has roughly 5% missing vales, and z_3 has 20% missing values. Altogether, about 44% subjects has at least one missing value in their covariates. The missing data mechanism is generated as MAR.

Missing values are filled in by using MICE. Ten multiple imputations are conducted in bootstrap stepwise procedures with 100 bootstrap replications. z_1 is forced in each of the best model by design in order to avoid a program crash while using the "step" function in R. Strong factors ($PI > 70\%$) and weak factors ($30\% < PI \leq 70\%$) are verified in the procedures based on percentage of inclusions (PI). Cox models are fitted using strong factors and the combination of weak factors. WBS's are calculated, and the model with smallest WBS is the best model. Traditional stepwise selection with fully observed data is conducted, and the results are compared with those of the proposed strategy.

7.1.1 No censoring in the outcomes

we assumed that there was no censoring in the data. Sample sizes were chosen to be 100 and 1000. 500 replications were performed for each sample size.

Sample Size 100 In the bootstrap stepwise procedures, the average PI of z_2 was 51.6%, and that of z_3 was 51.8%. z_2 was selected as a weak factor with 44.6% of the time and a strong factor with 28.4% of the time. z_3 was selected as a weak factor with 50.0% and a strong factor with 25.8%. The comparison among three variables are listed in Table 16, where 'PM' indicates the percentage of missingness in each variable, 'R' indicates the relationship with the outcome.

Table 12: A Comparison between Stepwise Selection for Fully Observed Data (F-SS) and the Proposed Stepwise Selection Strategy (P-SS) for Incomplete Data. (No censoring, and sample size 100 with 500 replications).

Variables	PM(%)	R	PI(%)	F-SS	P-SS
z_1	21	Strong	100	500	500
z_2	5	Weak	51.6	245	183
z_3	20	Weak	51.8	281	177

Sample Size 1000 The average PI of z_2 and z_3 were 97.0% and 98.3%, respectively. z_2 was selected as a weak factor 7 out of 500 times (1.4%) and a strong factor 493 out of 500 (98.6%) times. z_3 was selected as a weak factor 2 out of 500 times (0.4%)and a strong factor 498 out of 500 times (99.6%). The comparison among three variables are listed in Table 17.

7.1.2 Highly censored outcome data

We assumed 45% censoring in the event time. Sample sizes were chosen to be 100 and 1000 as well, and 500 replications were performed for each sample size.

Table 13: A Comparison between Stepwise Selection for Fully Observed Data (F-SS) and the Proposed Stepwise Selection Strategy (P-SS) for Incomplete Data. (No censoring, and sample size 1000 with 500 replications).

Variables	PM(%)	R	PI(%)	F-SS	P-SS
z_1	21	Strong	100	500	500
z_2	5	Weak	97.0	498	135
z_3	20	Weak	98.3	499	176

Sample Size 100 In the bootstrap stepwise procedures, the average PI of z_2 and z_3 were 43.1% and 44.3%, respectively. z_2 was selected as a weak factor 233 out of 500 times (46.7%) and a strong factor 76 out of 500 (15.2%) times. z_3 was selected as a weak factor 280 out of 500 times (56.0%) and a strong factor 68 out of 500 times (13.6%). The comparison among three variables are listed in Table 18.

Table 14: A Comparison between Stepwise Selection for Fully Observed Data (F-SS) and the Proposed Stepwise Selection Strategy (P-SS) for Incomplete Data. (45% censoring, and sample size 100 with 500 replications).

Variables	PM(%)	R	PI(%)	F-SS	P-SS
z_1	21	Strong	100	498	500
z_2	5	Weak	43.1	162	208
z_3	20	Weak	44.3	261	226

Sample Size 1000 The average PI of z_2 and z_3 were 89.1% and 88.7%, respectively. z_2 was selected as a weak factor 53 out of 500 times (10.6%) and a strong factor 445 out of 500 (89.0%) times. z_3 was selected as a weak factor 47 out of 500 times (9.4%)and a strong factor 450 out of 500 times (90.0%). The comparison among three variables are listed in

Table 19.

Table 15: A Comparison between Stepwise Selection for Fully Observed Data (F-SS) and the Proposed Stepwise Selection Strategy (P-SS) for Incomplete Data. (45% censoring, and sample size 1000 with 500 replications).

Variables	PM(%)	R	PI(%)	F-SS	P-SS
z_1	21	Strong	100	500	500
z_2	5	Weak	89.1	476	132
z_3	20	Weak	88.7	491	168

7.2 ALL CONTINUOUS COVARIATES I

Different simulation studies were conducted by generating all three covariates from normal distribtions. z_1 and z_2 were generated the same way with section 7.1. z_3 is a continuous variable from normal distribution $N(2, 36)$ and set up as weak factors by the strategy $z_3 = z_3 + t^3 + 2 * e_3$, where $e_3 \sim N(2, 100)$.

7.2.1 No censoring in the outcomes

We assumed that there is no censoring in the data. Simulations are conducted for sample sizes of 100 and 1000 with 500 replications.

Sample Size 100 In the bootstrap stepwise procedures, the average PI of z_2 was 53.4%, and that of z_3 was 37.7%. z_2 was selected as a weak factor 48.0% of the time and a strong factor 29.4% of the time. z_3 was selected as a weak factor 47.6% of the time and a strong factor 7.0%. The comparisons among the three variables are listed in Table 16.

Sample Size 1000 The average PI of z_2 and z_3 were 97.7% and 60.0%, respectively. z_2 was selected as a weak factor 7 out of 500 times (1.4%) and a strong factor 493 out of 500

Table 16: A Comparison between Stepwise Selection for Fully Observed Data (F-SS) and the Proposed Stepwise Selection Strategy (P-SS) for Incomplete Data. (No censoring, and sample size 100 with 500 replications).

Variables	PM(%)	R	PI(%)	F-SS	P-SS
z_1	21	Strong	100	500	500
z_2	5	Weak	53.4	242	163
z_3	20	Weak	37.7	123	190

(98.6%) times. z_3 was selected as a weak factor 2 out of 500 times (0.4%)and a strong factor 498 out of 500 times (99.6%). The comparison among three variables are listed in Table 17.

Table 17: A Comparison between Stepwise Selection for Fully Observed Data (F-SS) and the Proposed Stepwise Selection Strategy (P-SS) for Incomplete Data. (No censoring, and sample size 1000 with 500 replications).

Variables	PM(%)	R	PI(%)	F-SS	P-SS
z_1	21	Strong	100	500	500
z_2	5	Weak	97.7	498	135
z_3	20	Weak	98.3	499	176

7.2.2 Highly censored outcome data

We assumed 45% censoring in the event time. Sample sizes were chosen to be 100 and 1000 as well, and 500 replications were performed for each sample size.

Sample Size 100 In the bootstrap stepwise procedures, the average PI of z_2 and z_3 were 44.1% and 35.7%, respectively. z_2 was selected as a weak factor 239 out of 500 times (47.8%)

and a strong factor 83 out of 500 times (16.6%). z_3 was selected as a weak factor 227 out of 500 times (45.4%) and a strong factor 31 out of 500 times (6.2%). The comparisons among three variables are listed in Table 18.

Table 18: A Comparison between Stepwise Selection for Fully Observed Data (F-SS) and the Proposed Stepwise Selection Strategy (P-SS) for Incomplete Data. (45% censoring, and sample size 100 with 500 replications).

Variables	PM(%)	R	PI(%)	F-SS	P-SS
z_1	21	Strong	100	498	500
z_2	5	Weak	44.1	174	202
z_3	20	Weak	35.7	102	236

Sample Size 1000 The average PI of z_2 and z_3 were 89.1% and 88.7%, respectively. z_2 was selected as a weak factor 53 out of 500 times (10.6%) and a strong factor 445 out of 500 (89.0%) times. z_3 was selected as a weak factor 47 out of 500 times (9.4%) and a strong factor 450 out of 500 times (90.0%). The comparison among three variables are listed in Table 19.

Table 19: A Comparison between Stepwise Selection for Fully Observed Data (F-SS) and the Proposed Stepwise Selection Strategy (P-SS) for Incomplete Data. (45% censoring, and sample size 1000 with 500 replications).

Variables	PM(%)	R	PI(%)	T-SS	P-SS
z_1	21	Strong	100	500	500
z_2	5	Weak	89.1	498	131
z_3	20	Weak	88.7	358	191

7.3 ALL CONTINUOUS COVARIATES II

We generated z_1 the same way as in the last section. z_2 and z_3 were generated independently from the same normal distribution, $N(3, 25)$, and were set up as weak factors by the strategies $z_2 = z_2 + t + e_2 - 3$ and $z_3 = z_3 + t + e_3 - 3$, where $e_2 \sim N(2, 4)$ and $e_3 \sim N(2, 4)$ respectively.

7.3.1 No censoring in the outcomes

We assumed that there is no censoring in the data. Sample sizes were chosen to be 100 and 1000, and 500 replications were performed for each sample size.

Sample Size 100 In the bootstrap stepwise procedures, the average PI of z_2 was 49.7%, and that of z3 was 48.1%. z_2 was selected as a weak factor with 49.0% of the time and a strong factor with 23.4% of the time. z_3 was selected as a weak factor with 53.2% and a strong factor with 19.4%. The comparisons among three variables are listed in Table 20,

Table 20: A Comparison between Stepwise Selection for Fully Observed Data (F-SS) and the Proposed Stepwise Selection Strategy (P-SS) for Incomplete Data. (No censoring, and sample size 100 with 500 replications).

Variables	PM(%)	R	PI(%)	F-SS	P-SS
z_1	21	Strong	100	500	500
z_2	5	Weak	49.7	229	166
z_3	20	Weak	48.1	234	166

Sample Size 1000 The average PI of z_2 and z_3 were 97.7% and 60.0%, respectively. z_2 was selected as a weak factor 7 out of 500 times (1.4%) and a strong factor 493 out of 500 (98.6%) times. z_3 was selected as a weak factor 2 out of 500 times (0.4%) and a strong factor 498 out of 500 times (99.6%). The comparisons are listed in Table 21.

Table 21: A Comparison between Stepwise Selection for Fully Observed Data (F-SS) and the Proposed Stepwise Selection Strategy (P-SS) for Incomplete Data. (No censoring, and sample size 1000 with 500 replications).

Variables	PM(%)	R	PI(%)	F-SS	P-SS
z_1	21	Strong	100	500	500
z_2	5	Weak	97.5	498	140
z_3	20	Weak	92.4	499	133

7.3.2 Highly censored outcome data

We assumed 45.0% censoring in the event time. Sample sizes were chosen to be 100 and 1000 as well, and 500 replications were performed for each sample size.

Sample Size 100 In the bootstrap stepwise procedures, the average PI of z_2 and z_3 were 44.1% and 35.7%, respectively. z_2 was selected as a weak factor 239 out of 500 times (47.8%) and a strong factor 83 out of 500 (16.6%) times. z_3 was selected as a weak factor 227 out of 500 times (45.4%) and a strong factor 31 out of 500 times (6.2%). Comparisons among three variables are listed in Table 22.

Table 22: A Comparison between Stepwise Selection for Fully Observed Data (F-SS) and the Proposed Stepwise Selection Strategy (P-SS) for Incomplete Data. (45.0% censoring, and sample size 100 with 500 replications).

Variables	PM(%)	R	PI(%)	F-SS	P-SS
z_1	21	Strong	100	498	500
z_2	5	Weak	43.6	164	205
z_3	20	Weak	42.4	189	234

Sample Size 1000 The average PI of z_2 and z_3 in the bootstrap stepwise procedures were 89.2% and 81.5%, respectively. z_2 was selected as a weak factor 45 out of 500 times (8.6%) and a strong factor 455 out of 500 (91.0%) times. z_3 was selected as a weak factor 100 out of 500 times (20.0%) and a strong factor 394 out of 500 times (78.8%). Comparisons are listed in the Table. 23.

Table 23: A Comparison between Stepwise Selection for Fully Observed Data (F-SS) and the Proposed Stepwise Selection Strategy (P-SS) for Incomplete Data. (45.0% censoring, and sample size 1000 with 500 replications).

Variables	PM(%)	R	PI(%)	F-SS	P-SS
z_1	21	Strong	100	500	500
z_2	5	Weak	89.2	479	126
z_3	20	Weak	81.5	482	156

7.4 APPLICATION TO NSABP DATA

The NSABP data was described earlier in sections 1.2 and 3.6.4. As was stated earlier, 49.52% of the patients had at least one missing values in the covariates. We first conduct 10 multiple imputations to fill-in missing values in the imputation step. Then, we generated $B = 100$ bootstrap samples of the same size [51] with the complete data sets. For the NSABP data, the results of bootstrap stepwise selection are presented in Table 24. The strong factors are age^2, tumor type, nuclear grade, nodal status and race; the weak factors are age, blood vessel invasion and ER. The PI of PR is only 27.3%, so it will be omitted from the covariate pool that will be used for the weak factor selection step. A similar conclusion can be made when using $B = 500$ bootstrap replications (Table 25). Since age and age^2 are terms describing the relationship age at entry with survival, we would want both to be included in models for weak factor selection.

Table 24: The PI for each covariate in the bootstrap stepwise selections $B = 100$

Variables	Age	Age^2	blood vessel invasion	ER	Tumor type1*
PI (%)	50.5	99.6	62.6	47.8	87.9
Variables	Tumor type2*	Nuclear grade	Nodal status	PR	Race
PI (%)	97.5	98.8	100.0	27.3	87.0

$*$ indicates two dummy variables for the tumor type

Table 25: The PI for each covariate in the bootstrap stepwise selections $B = 500$

Variables	Age	Age^2	blood vessel invasion	ER	Tumor type1*
PI (%)	49.5	99.5	62.0	45.9	88.4
Variables	Tumor type2*	Nuclear grade	Nodal status	PR	Race
PI (%)	97.8	98.7	100.0	26.9	86.5

$*$ indicates two dummy variables for the tumor type

In the NSABP data, age and nodal status were fully observed, and the other variables had at least one missing value. Therefore, the model for π_i can be written as

$$\hat{\pi}_i = Pr(r_i = 1|T_i, \delta_i, z_{i,age}, z_{i,age^2}, z_{i,node}) = \frac{\exp(-\mathbf{v}'\mathbf{m}_i)}{1 + \exp(-\mathbf{v}'\mathbf{m}_i)} , \tag{7.1}$$

and

$$\mathbf{v}'\mathbf{m}_i = v_0 + v_1 * t_i + v_2 * \delta_i + v_3 * z_{i,age} + v_4 * z_{i,age^2} + v_5 * z_{i,node} . \tag{7.2}$$

A logistic regression was used to model the function of π_i as in (7.1). In order to avoid misspecification of π_i, we need to fit π_i as well as possible. The variables whose significant level less than and equal to 0.2 will stay in the model [60]. Therefore, age and age^2 are discarded from model. Hence, the model for π_i is given by

$$\hat{\pi}_i = Pr(r_i = 1|T_i, \delta_i, z_{i,node}) = \frac{\exp[-(v_0 + v_1 * t_i + v_2 * \delta_i + v_3 * z_{i,node})]}{1 + \exp[-(v_0 + v_1 * t_i + v_2 * \delta_i + v_3 * z_{i,node})]} . \tag{7.3}$$

Cox models were fitted by using strong factors and the combination of weak factors. The WBS's for models can be calculated based on the $\hat{\pi}_i$ (Equation 7.3). In the bootstrap stepwise procedure, strong factors were square of age, race, nodal status, nuclear grade and tumor type, and weak factors were estrogen receptor and blood vessel invasion. Therefore, four models are fitted according to the model selection strategy:

- model 1: COX= $age + age^2 + race + node + nuclear.grade + tumor.type + blood + er$.
- model 2: COX= $age + age^2 + race + node + nuclear.grade + tumor.type + blood$.
- model 3: COX= $age + age^2 + race + node + nuclear.grade + tumor.type + er$.
- model 4: COX= $age + age^2 + race + node + nuclear.grade + tumor.type$.

Table 26 lists the result of WBS in each model. Model 4's WBS is the smallest one (0.1727), which indicates that there is no gain in the predictive accuracy of model fitting after weak factors are entering the model. Therefore, all the weak factors should be out. The final model of the model selection is model 4.

Table 26: The weighted Brier Score for four models

Model	1	2	3	4
WBS	0.1880	0.1848	0.1751	0.1727

7.5 DISCUSSION

Our simulation results indicate that the amount of missing data does not affect strong factors in model selections, but it does affect weak factors. We found that the more missing values that are in the weak factors, the less chance that the factors are selected into the final model. A possible reason is that the use of imputations increases the standard errors.

When a sample size is large, the proposed strategy using WBS tends to exclude weak factors from the final models as indicated by the low numbers of times that weak factors

are selected into the final models. We believe that there are two reasons for this. First, the imputations for large sample sizes generate more noise (large standard errors), which are sensitive to the evaluation of the predictive accuracy of potential models in the proposed strategy. Second, the bounds for the inclusion criteria ($30\% < PI \leq 70\%$) are arbitrary. The upper bound appears to be reasonable as strong factors tend to be included frequently as we would want. The lower bound of 30% PI for weak factors may result in a lower percentage of weak factors ultimately being included in the final models.

If the desire is to have weak factors frequently included in the fianl model, then, perhaps, adjusting the lower PI bound upwards, say to 50%, would accomplish this. In most traditional stepwise procedures, once the sample size is very large, all factors, strong or weak, are guaranteed to be included in a final model. For weak factors, even a small effect would result in the factor being included. Some authors (e.g., Kass and Raftery [61]) argue that in some cases, evidence supports the null hypothesis, i.e., the exclusion of weak factors in our case. It may be that our strategy builds in a method of exclusion of weak factors even when a sample size is large.

A possible extension of this work is that $\hat{\pi}_i$ in the WSF estimator can be smoothed using a kernel smoother [62]. An example of such a smoother is the Nadaraya–Watson estimator [63] which is represented as

$$\hat{\pi}_i(\mathbf{w}) = \frac{\sum_{i=1}^{n} r_i K_h(\mathbf{w} - \mathbf{W}_i)}{\sum_{j=1}^{n} K_h(\mathbf{w} - \mathbf{W}_j)} \ , \tag{7.4}$$

where $\mathbf{W}$ is defined as above, K is a d^{th} order kernel function of $\mathbf{W}$ satisfying $\int K(u)du = 1$, $\int u^m K(u)du = 0$ for $m = 1, \ldots, (d-1)$, $\int u^d K(u)du \neq 0$, and $\int K(u)^2 du < \infty$. Also, $K_h(\cdot) = K(\cdot/h)$, where h is the smoothing parameter or the *bandwidth*.

Variable selection and related issues have been investigated for years, but all of the available methods are based on fully observed data or complete cases analysis. However, when there are missing values in the covariates, complete cases analysis would give unreasonable model selection results. In this dissertation, the proposed three–step model selection strategy takes into account information due to missing values by defining a WBS, which provides an alternative to using p–values in the model selection process. In particular, our method accommodates lost information due to unobserved data during the stepwise procedure, how

to include or exclude weak factors in a final model, and can be used to determine whether or not the amount of missing values or censoring affects the model selection results. Based on our simulation studies, our present recommendations for a model selection strategy are to only use bootstrap stepwise procedures to select covariates when a sample is large, and to use the proposed model selection strategy when a sample is small.

On the other hand, there are some unresolved questions about the proposed three-step strategy based on the simulation studies. For example, in highly censored outcome data, the proposed strategy tends to include weak factors in final models when sample size is 100, which is a different result compared to other simulation results. Moreover, for large sample sizes, weak factors were selected into final models lower number of times compared with small sample sizes, which is odd, because it tends to select factors more times for large sample sizes than small sample sizes. However, because our strategy involves the use of predictive accuracy, it may be that "weak" factors are only included in a final model some fixed proportion of the time. If that is the case, that proportion may not increase with an increasing sample size. These questions are interesting and worthy for further reasearch.

APPENDIX A

PROGRAM TO COMPARE MICE AND MIX

```
for (k in 1:NT){
z1<-rnorm(N,a1[1],a1[2])         #<--distribution of z1~N(a10,a11)
z2<-rnorm(N,a2[1],a2[3])         #<--distribution of z2~N(a21,a22)
z3<-rnorm(N,a3[3],a3[4])
z4<-rbinom(N,1,0.6)
p<-exp(aa[1]+aa[2]*z4)/(1+exp(aa[1]+aa[2]*z4))
z5<-rbinom(N,1,p)
#<-- z5 is from Bernoulli(p) and z5 can be missing for some subjects
z5[z5==0]<-2       #<-- the binary in mix should be 1 and 2, not 1 and 0.
z4[z4==0]<-2

thr <- exp(beta[1]*z1+beta[2]*z2+beta[3]*z3+beta[4]*z4+beta[5]*z5)
#<-- True exponential hazards

prop.cens<-0.1
cen<-rbinom(N,1,1-prop.cens) #<--censor indicator:10% censor in the case.
t<- rexp(N)/thr

#<----------------------fully observed case----------------------
all<-data.frame(z4,z5,z1,z2,z3)
fit.model.all<- coxph(Surv(t,cen) ~ z4 + z5+z1+z2+z3,data=all)
out1[, k] <- c(fit.model.all$coef, diag(fit.model.all$var))

#<-------------------------------------------------------------------
#<----------------------generate missing in the covariates-----------

p.z2mis<-(exp(phi1[1]+phi1[2]*t+phi1[3]*z1))
            /(1+exp(phi1[1]+phi1[2]*t+phi1[3]*z1))
p.z3mis<-((phi2[1]+phi2[2]*t+phi2[3]*z1+phi2[4]*z2))
           /(1+(phi2[1]+phi2[2]*t+phi2[3]*z1+phi2[4]*z2))

#----------generate missings for z2 & z3
uu<-runif(N)
r2<-rep(0,N)
r3<-rep(0,N)
r2[uu<(1-p.z2mis)]<-1
```

```
r3[uu<(1-p.z3mis)]<-1
z2[r2==0]<-NA
z3[r3==0]<-NA
z2.1<-rep(0,N)
z3.1<-rep(0,N)
z5.1<-rep(0,N)

z2.1<-z2
z3.1<-z3
z5.1<-z5
#----------generate missings for z5
u<-mean(t)/sqrt(var(t))  #<-- x* in the missing data mechaniam
pp<-exp(fai[1]+fai[2]*u+fai[3]*z4+fai[4]*z4*u)
       /(1+ exp(fai[1]+fai[2]*u+fai[3]*z4+fai[4]*z4*u))
rrf<-rbinom(N,1,pp)
#<-- the distribution of rr(missing data mechaniam) is bernoulli
z5[rrf==0]<-NA
#<----------------cox model by using complete case--------------------
cc<-data.frame(z4,z5,z1,z2,z3)
fit.model.cc<-coxph(Surv(t,cen)~z4+z5+z1+z2+z3,na.action=na.exclude,data=cc)
out2[, k] <- c(fit.model.cc$coef, diag(fit.model.cc$var))

#------------make a matrix of covariates via mice----------------------#
Y<-data.frame(z4,z5,z1,z2,z3)         #<--Y has to be a matrix or dataframe
mm<-10         #<--number of imputation
imp<-mice(Y,m=10,seed=123456+k)
#imp<-mice(Y,m=10)
#mi<-matrix()
mi<-data.frame()
coef.mi<-matrix(ncol=mm,nrow=5)       #<--rows are covariates and col are mi
coef.mi.var<-matrix(ncol=mm,nrow=5)

for (j in 1:mm){
mi<-(complete(imp,j))       #<--completed data sets after MI
fit.model.mi<-coxph(Surv(t,cen)~z4+z5+z1+z2+z3,data=mi)
coef.mi[,j]<-fit.model.mi$coef
coef.mi.var[,j]<-diag(fit.model.mi$var)
}

coef.mi.tot<-list(coef.mi[1,],coef.mi[2,],coef.mi[3,]
             ,coef.mi[4,],coef.mi[5,])
mi.coef<-sapply(coef.mi.tot,mean)
#<--means of coef of MI
mi.bv<-sapply(coef.mi.tot,var)
#<--between var
var.mi.tot<-list(coef.mi.var[1,],coef.mi.var[2,],
            coef.mi.var[3,],coef.mi.var[4,],coef.mi.var[5,])
mi.v<-sapply(var.mi.tot,mean)         #<--within var
mi.total<-mi.v+(1+(1/mm))*mi.bv       #<--total var of MI

dfm<-(mm-1)*((1+1/(mm+1))*mi.v/mi.bv)^2
ci.low<-mi.coef- qt(.975, df = dfm)*sqrt(mi.total)
ci.up<-mi.coef+ qt(.975, df = dfm)*sqrt(mi.total)
hat.im[[k]]<-c(data.frame(mi.coef,mi.total,ci.low,ci.up))
```

```
#------------make a matrix of covariates-----------------------#
Y1=cbind(z4,z5.1,z1,z2.1,z3.1)  #<--Y has to be a matrix for mix
s1<-prelim.mix(Y1,2)

#---------------------------run MI via mix------------------------------#

MI<-vector("list",10)  #<--vector of complete data after MI
fit.model.mi<-vector("list",10)
rngseed(1234567)       #<-- set random number generator seed


for (i in 1:10){

     cat("Doing imputation ",i,"\n")
       thetahat <- em.mix(s1)
       newtheta <- da.mix(s1, thetahat, steps=500, showits=TRUE)
       MI[[i]] <- imp.mix(s1, newtheta)
     }


M10<-coxph(Surv(t,cen)~MI[[10]][,1]+MI[[10]][,2]+MI[[10]][,3]
                       +MI[[10]][,4]+MI[[10]][,5])
M9<-coxph(Surv(t,cen)~MI[[9]][,1]+MI[[9]][,2]+MI[[9]][,3]+MI[[9]][,4]
                       +MI[[9]][,5])
M8<-coxph(Surv(t,cen)~MI[[8]][,1]+MI[[8]][,2]+MI[[8]][,3]+MI[[8]][,4]
                       +MI[[8]][,5])
M7<-coxph(Surv(t,cen)~MI[[7]][,1]+MI[[7]][,2]+MI[[7]][,3]+MI[[7]][,4]
                       +MI[[7]][,5])
M6<-coxph(Surv(t,cen)~MI[[6]][,1]+MI[[6]][,2]+MI[[6]][,3]+MI[[6]][,4]
                       +MI[[6]][,5])
M5<-coxph(Surv(t,cen)~MI[[5]][,1]+MI[[5]][,2]+MI[[5]][,3]+MI[[5]][,4]
                       +MI[[5]][,5])
M4<-coxph(Surv(t,cen)~MI[[4]][,1]+MI[[4]][,2]+MI[[4]][,3]+MI[[4]][,4]
                       +MI[[4]][,5])
M3<-coxph(Surv(t,cen)~MI[[3]][,1]+MI[[3]][,2]+MI[[3]][,3]+MI[[3]][,4]
                       +MI[[3]][,5])
M2<-coxph(Surv(t,cen)~MI[[2]][,1]+MI[[2]][,2]+MI[[2]][,3]+MI[[2]][,4]
                       +MI[[2]][,5])
M1<-coxph(Surv(t,cen)~MI[[1]][,1]+MI[[1]][,2]+MI[[1]][,3]+MI[[1]][,4]
                       +MI[[1]][,5])


est.coef<-list(M1$coef,M2$coef,M3$coef,M4$coef,M5$coef,
               M6$coef,M7$coef,M8$coef,M9$coef,M10$coef)
est.st<-list(sqrt(diag(M1$var)),sqrt(diag(M2$var)),sqrt(diag(M3$var)),
             sqrt(diag(M4$var)),sqrt(diag(M5$var)),sqrt(diag(M6$var)),
             sqrt(diag(M7$var)),sqrt(diag(M8$var)),sqrt(diag(M9$var)),
             sqrt(diag(M10$var)))
hat.im1[[k]]<-mi.inference(est.coef,est.st,confidence=0.95)
}
```

APPENDIX B

PROGRAM FOR THE PROPOSED MODEL SELECTION STRATEGY

```
#########################################################
# No censoring in the data
# z1-strong and large missing,
# z2-weak and small missing,
# z3-weak and large missing.
#########################################################
for (k in 1:NT){

z1<-rnorm(N,0,1)
z2<-rnorm(N,3,5)
z3<-rbinom(N,1,0.6)

t<- rexp(N)
cen<-rep(1,N)
#prop.cens<-0.45
#cen<-rbinom(N,1,1-prop.cens)
#t<- rexp(N)


e<-rnorm(N)
z1<-t+z1+e    #<--z1 has strong relationship with outcome

e2<-rnorm(N,2,2)
z2<-z2+t+e2-3

e3<-rnorm(N,2,10)
z3<-z3+t^3+2*e3
z3[z3>0]<-1
z3[z3<=0]<-0

all.data<-data.frame(z1,z2,z3)
fit.model.all<- coxph(Surv(t,cen) ~z1+z2+z3,data=all.data)
#fit.model.all
reg.model.all<-step(fit.model.all)
select.v[[k]]<-reg.model.all$coef

#----------------------Generate missing values----------------------#
```

```
z1.m<-rbinom(N,1,0.79)
z1[z1.m==0]<-NA

z2.m<-rbinom(N,1,0.95)
z2[z2.m==0]<-NA

z3.m<-rbinom(N,1,0.8)
z3[z3.m==0]<-NA

#--------------------fix Cox model by CC---------------------------------#
cc<-data.frame(z1,z2,z3)
fit.model.cc<-coxph(Surv(t,cen)~z1+z2+z3,na.action=na.exclude,data=cc)
fit.model.cc

#------------make a matrix of covariates via mice-----------------------#
Y<-data.frame(z1,z2,z3)         #<--Y has to be a matrix or dataframe
mm<-10         #<--number of imputation
imp<-mice(Y,m=10,seed=123456+k)

##-------------------------------------------------------------------------##
##- conduct bootstrap-stepwise produres in each imputed data set;      ##
##- there are 10 imputations, and each has B bootstrap samples;        ##
##- stepwise procedures are carried out to select variables at each    ##
##- bootstrap samples. Therefore if a variable were selected at every  ##
##- bootstrap smaples, the counts of the variable included are 10*B.   ##
##-------------------------------------------------------------------------##

B<-100
reg.bs<-matrix(nrow=B,ncol=3) #<--nrow= # of bootstrap,ncol= # of z'col
reg.step<-list()              #<--list of coef of chosen model in stepwise
chosen.vb<-list()
tot.vb<-list()
vb<-vector()

for (j in 1:mm){

  for (b in 1:nrow(reg.bs)){           #<--bootstrap 10 times
  mi<-data.frame(t,cen,complete(imp,j))#<--completed data sets after MI
  in.star<-sample(1:N,N,replace=T)     #<--to sample id with replace.
  data.star<-mi[in.star,]              #<--create a bootstrap sample

  M<-coxph(Surv(data.star$t,data.star$cen)~
           data.star$z1+data.star$z2+data.star$z3,data=data.star)

  reg.model<-stepAIC(M,scope = list(lower = ~data.star$z1))
  #<--carry out stepwise procedures
  reg.step[[b]]<-c((reg.model$coef))
  #<--variables are selected into final models at each bootstrap sample.
  chosen.vb[[b]]<-names(reg.step[[b]])
  #<--record names variables included in final models
}
tot.vb[[j]]<-c(unlist(chosen.vb))
}
```

```
variables<-unlist(tot.vb)               #<--change to a vector
poi<-table(variables)
out1[,k]<-c(poi)                        #<--perctage of inclusion

#########################################################################
#                    A model selection strategy                         #
#########################################################################

Y.variables<-data.frame(z1,z2,z3)#<--only include the variables with missing
indi<-apply(Y.variables,1,mean)
r<-rep(0,N)
r[!is.na(indi)]<-1 #<--r=0 if indi=NA. An indicator of missing covariates

#<-----------------------to fit model for pi---------------
dpi<-data.frame(r,t,cen)
fit.pi<-glm(r~t+cen,data=dpi,family=binomial)
#<--only t and cen are fully obs
log.odds<-predict.glm(fit.pi,type = "response")
#<--the default predictions are of log-odds
#<--(probabilities on logit scale log(pi/(1-pi))
pi<-log.odds/(1+log.odds)
#-------------------to compute weighted baseline hazard-----
#-------------------------------------------------------------
#<--use complete data set mi in bootstrap stepwise iteration
#-------------------------------------------------------------
fit.model1<-coxph(Surv(mi$t,mi$cen)~mi$z1+mi$z2+mi$z3,data=mi)
bz.obs<-fit.model1$coef[2]*mi$z2+fit.model1$coef[1]*mi$z1
        +fit.model1$coef[3]*mi$z3
bz.imp<-fit.model1$coef[1]*mi$z1+fit.model1$coef[2]*mi$z2
        +fit.model1$coef[3]*mi$z3
wimp<-(1-r/pi)*exp(bz.imp)
wimp[wimp<0]<-0
#<--if the elements of z1, z3 are obs, then (1-r/pi)*exp(bz.imp)=0
#<--because obs contributed to (r/pi)*exp(bz.obs)
wi<-(r/pi)*exp(bz.obs)+wimp
#<--there should be a sum based on subjects still at risk on time t_j.

w.data<-data.frame(mi,r,pi,wi)
#<--r,pi and weight of each subject are added to new data.
i<-(order(w.data[,1]))
#<--i<-(order(w.data[,1]))ascending order
for.ss<-w.data[i,]

i<-rev(order(w.data[,1]))#<--i<-(order(w.data[,1]))ascending order
for.s<-w.data[i,]


#-----------------------compute w and s----------------------------
ww<-cumsum(for.ss$wi)
wt<-cbind(for.ss$wi,ww)
ss<-(1-for.ss$cen/ww)
s0<-cumprod(abs(ss))   #<--baseline survival
```

```
#----------------------------compute G(T) in the Breir score----------
G<-survfit(Surv(mi$t,mi$cen)~1,type=("kaplan-meier"),data=mi)
G.t<-summary(G) #<--n.risk=894

#----------------------------compute s0 and s-----------------------
nocen<-cbind(for.ss$t,for.ss$cen, for.ss$wi,s0)
event.table<-cbind(G.t$time,nocen[match(G.t$time,nocen[,1]),2],
    nocen[match(G.t$time,nocen[,1]),3],nocen[match(G.t$time,nocen[,1]),4])
#<--censor and wi are matched with event time according to ff$time
#<--but this 'match' deleted all duplicated event time.
#<-- the smaller event time is, the larger sum of w is and the larger s is.
s<-event.table[,4]^abs((event.table[,3]))
j<-rev(order(s[]))
s<-s[j]
life.table<-cbind(G.t$time,G.t$surv,s)

##-------------------------------------------------------------------------
##--fit Cox models with strong factors and combination of weak factors
##-------------------------------------------------------------------------
summary(t)
ins<-G.t$surv
G.t$surv[ins==0]<-0.00001
cat1<-cumsum(s^2*(1/G.t$surv))

#<--BS=>
#bs<-(cat1[49]+cat3[1])/N
bs1<-cat1[length(G.t$time)]/N
```

BIBLIOGRAPHY

[1] Cox, D. R. Regression models and life-tables. (with discussion). *Journal of the Royal Statistical Society B*, 34:187-220, 1972.

[2] Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71: 431-444, 1984.

[3] Miller, A. J. Selection of subsets of regression variables (with discussion). *Journal of the Royal Statistical Society, Series A*, 147:389-425, 1984.

[4] Little, R. J., and Rubin, D. B. *Statistical analysis with missing data. Second edition.* Wiley & Son:Hoboken, New Jersey, 2002.

[5] Lipsitz, S. R., Ibrahim, J. G. and Zhao, L. P. A weighted estimating equation for missing covariate data with propertied similar to maximum likelihood. *Journal of the American Statistical Association*, 94:1147-1160, 1999.

[6] Wang, C.Y. ,and Chen, H. Y. Augmented inverse probability weighted estimator for Cox missing covariate regression. *Biometrics*, 57:414-419, 2001.

[7] Parzen, M., Lipsitz, S. R., Ibrahim, J. G., and Lipshultz, S. A weighted estimating equation for linear regression with missing covariate data. *Statistics in Medicine*, 21:2421-2436, 2002.

[8] Lin, D. ,and Ying, Z. Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, 88:1341-1349, 1993.

[9] Lipsitz, S.R., and Ibrahim, J. G. Estimating Equations with incomplete categorical covariates in the cox model. *Biometrics*, 54:1002-1123, 1998.

[10] Herring, A., and Ibrahim, J. Likelihood-based methods for missing covariates in the Cox proportional hazarded model. *Journal of the American Statistical Association*, 96:292-302, 2001.

[11] Paik, M. C., and Tsai, W. Y. On using the Cox proportional hazards model with missing covariates. *Biometrika*, 84:579-593, 1997.

[12] Buuren, S. V., Boshuizen, H. C., and Knook, D. L. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18:681-694, 1999.

[13] Harrell, F. E. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer:New York, 2001.

[14] Fisher, B., Anderson, S., Bryant, J., Margolese, R. G., Deutsch, M., Fisher, E. R., Jeong, J., and Wolmark, N. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancar. *The New England Joural of Medicine*, 347:1233-1241, 2002.

[15] Fisher, E. R., Anderson, S., Tan-Chiu, E., Fisher, B., and Wolmark, N. Fifteen-year prognostic discriminants for invasive breast carcinoma. *Cancer*, 91:1679-1687, 2001.

[16] Robins, J. M., Rotnitzky, A. ,and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846-866, 1994.

[17] Henderson, H. V., and Velleman, P. F. Building multiple regression models interactively. *Biometrics*, 37:391-411, 1981.

[18] Schafer, J. L. *Analysis of imcomplete multivariate data*. Chapmen and Hall: New York, 1997.

[19] Sauerbrei, W., and Schumacher, M. A Bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine*, 11:2093-2109, 1992.

[20] Robin, D. B. *Multiple imputations in sample surveys-A phenomenological Bayesian approach to nonresponse*. Imputation and Editing of Faulty or Missing Survey Data, US Department of commerce, 1978.

[21] Paik, M. C. Multiple imputation for the Cox proportional hazards model with missing covariates. *Lifetime Data Analysis*, 3:289-298, 1997.

[22] Cho, M., and Schenker,N. Fitting the log-F accelerated failure time model with incomplete covariate data. *Biometrics*, 55:826-833, 1999.

[23] Sartori, N., Salvan, A., and Thomaseth, K. Multiple imputation of missing values in a cancer mortality analysis with estimated expose dose. Preprint submitted to *Elsevier Science*, 2004.

[24] Wang, C. Y., Wang, S., Zhao, L., and Ou, S. Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, 92:512-524, 1997.

[25] Zhou, H., and Pepe, M. S. Auxiliary covariate data in failure tie regression. *Biometrika*, 82:129-149, 1995.

[26] Lipsitz, S.R., and Ibrahim, J. G. Using the EM algorithm for survival data with incomplete categorical covariates. *Lifetime Date Analysis*, 2:5-14,1996.

[27] Chen, H. Y., and Little, R. L. Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 94:896-908, 1999.

[28] Ibrahim, J. G., Lipsitz, S. R., and Chen, M. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society B*, 61:173-190,1999.

[29] Gelfand, A. E., and Smith, A. F. M. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398-409, 1990.

[30] Gilks, W. R., and Wild, P. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41:337-348, 1992.

[31] Olkin, I., and Tate, R. F. Multivariate correlation models with mixed discrete and continuous variables. *Annals of Methematical Statistics*, 32:448-465, 1961.

[32] Little, R. J. A., and Schluchter, M. D. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72:497-512, 1985.

[33] Tanner, M. A., and Wong, W. H. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82:528-550, 1987.

[34] Rubin, D. B. Inference and missing data. *Biometrika*, 63:581-592, 1976.

[35] Horton, N. J. and Lipsitz, S. R. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistican*, 55:244-254, 2001).

[36] Buuren, S.V. and Oudshoorn, G.G.M. *Multivariated imputation by chained equations. MICE V1.0 User's manual*. TNO Report PG/VGZ/00.038, Rotterdam/TNO Prevention and Health, Leiden, 2000.

[37] Brand, J.P.L. *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis*. Academic thesis, Erasmus University, Rotterdam/TNO Prevention and Health, Leiden, 1999.

[38] Olkin, I., and Tate, R. F. Multivariate correlation models with MIXed discrete and continuous variables. *Annals of Methematical Statistics*, 32:448-465, 1961.

[39] Burton, A. and Altman, D. G. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *British Journal of Cancer*, 91:4-8, 2004.

[40] Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. Survival analysis Part IV: Further concepts and methods in survival analysis. *British Journal of Cancer*, 89:781-786, 2003.

[41] Harrell, F. E., Lee K. L., and Mark, D. B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Mecidine*, 15:361-387, 1996.

[42] Steyerberg, E. W., Eijkemans, and Habbema, J. D. F. Stepwise selection in small data sets: A sitmulation study of bias in Logistic regression analysis. *Journal of Clinical Epidemiology*, 52:935-942, 2000.

[43] Wählby, U., Jonsson, E. N., and Karlsson, M. O. Comparison of stepwise covariate model building strategies in population pharmacokinetic-pharmacodynamic analysis *AAPS PharmSci*, 4:1-12, 2002.

[44] Ambler, G., Brady, A. R., and Royston, P. Simplifying a prognostic model: a simulation study based on clinical data. *Statistics in Medicine*, 21:3803-3822, 2002.

[45] Altman, D. G., and Andersen, P. K. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, 8:771-783, 1989.

[46] Austin, P. C., and Tu, J. V. Automated variable selection methods for logistic regression produced unstable models for prediction acute myocardial infarction mortality. *Journal of Clinical Epidemiology*, 57:1138-1146, 2004.

[47] Angustin, N., Sauerbrei, W., and Schumacher, M. The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistics Modeling*, 5:95-118, 2005.

[48] Chen, C., and George, S. L. The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Statistics in Medicine*, 4:39-46, 1985.

[49] Sauerbrei, W. The use of resampling methods to simplify regression models in medical statistics. *Apply Statistics*, 48:313-339, 1999.

[50] Efron, B. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 1:1-26, 1979.

[51] Schumacher, M., Holländer, N. and sauerbrei, W. Resampling and cross-validation techniques: a tool to reduce bias cased by model building? *Statistics in Medicine*, 16:2813-2827, 1997.

[52] Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18:2529-2545, 1999.

[53] Breiman, L. and Spector, P. Submodel selection and evaluation in regression. The X-random case. *International Statistical Review*, 60:291-319, 1992.

[54] Breiman, L. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24:2350-2383, 1996.

[55] Chung, H., Lee, K. and Koo, J. A note on bootstrap model selection criterion. *Statistics and Probability Letters*, 26:35-41, 1996.

[56] Zucchini, W. An introduction to model selection. *Journal of Mathematical Psychology*, 44:41-61, 2000.

[57] Schumacher, M and Gerds, E. G. How to assess prognostic models for survival data: a case study in oncology . *Statistics in Medicine*, 42:564-571, 2003.

[58] Rosthøj, s. and keiding, N. Explained variation and predictive accuracy in general parametric statistical models: the role of model misspecification. *Lifetime Data Analysis*, 10:461-472, 2004.

[59] Klein, J. P. and Moeschberger, M. L. *Survival analysis. Techniques for censored and truncated data*. Springer: New York, 2003.

[60] Parzen, M., Lipsitz, S. R., Ibrahim, J. G. and Lipshultz, S. A weighted estimating equation for linear regression with missing covariate data . *Statistics in Medicine* , 21:2421-2436, 2002.

[61] Kass, R. and Raftery, A. E. Bayes factors. *Journal of the American Statistical Association*, 90:773-795, 1995.

[62] Qi, L., Wang, C.Y. and Prentice, R.L. Weigted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 100:1250-1263, 2005.

[63] Simonoff, J.S. *Smoothing methods in statistics*. Springer: New York, 1996.